Journal of Artificial General Intelligence 6(1) 5-20, 2015
DOI: 10.1515/jagi-2015-0002

Submitted 2015-05-17
Accepted 2015-11-19

Unnatural Selection: Seeing Human Intelligence in Artificial Creations

Tony Veale TONY.VEALE@UCD.IE
School of Computer Science
University College Dublin
Belfield, Dublin D4, Ireland

Editor: Tarek R. Besold, Kai-Uwe Kühnberger, Tony Veale

Abstract

As generative AI systems grow in sophistication, so too do our expectations of their outputs. For as automated systems acculturate themselves to ever larger sets of inspiring human examples, the more we expect them to produce human-quality outputs, and the greater our disappointment when they fall short. While our generative systems must embody some sense of what constitutes human creativity if their efforts are to be valued as creative by human judges, computers are not human, and need not go so far as to actively pretend to be human to be seen as creative. As discomfiting objects that reside at the boundary of two seemingly disjoint categories, creative machines arouse our sense of the uncanny, or what Freud memorably called the *Unheimlich*. Like a ventriloquist's doll that finds its own voice, computers are free to blend the human and the non-human, to surprise us with their knowledge of our world and to discomfit with their detached, other-worldly perspectives on it. Nowhere is our embrace of the unnatural and the uncanny more evident than in the popularity of *Twitterbots*, automatic text generators on Twitter that are followed by humans precisely because they are non-human, and because their outputs so often seem meaningful yet unnatural. This paper evaluates a metaphor generator named *@MetaphorMagnet*, a Twitterbot that tempers the uncanny with aptness to yield results that are provocative but meaningful.

Keywords: computational creativity, language, art, readymades, modernism, Twitterbots

1. Introduction

Each century poses new challenges to society's concepts of art and creativity (Hughes, 1991). Foremost of the 20th century's challenges was Marcel Duchamp's concept of readymade art (Duchamp, 1917; Taylor, 2009). When Duchamp exhibited a signed urinal named *Fountain* at a Dadaist show in 1917, viewers were challenged to see art as more than a product of the artist's labors, but as the deliberate choice of an artist to select and frame objects in ways that add new meanings and new resonance. Yet for many at first, including the event's organizers (Kuenzli and Naumann, 1989), Duchamp's *Fountain* appeared to be a category error, a tawdry object drawn from a category of banal objects that should never be accepted as art.

Yet other artists would build on Duchamp's subversion of the familiar to pose further challenges to our received ideas about art, as when René Magritte highlighted the "treachery" of entrenched mappings from words or images to their habitual meanings. In his *Les Mots et Les Images* (Magritte, 1929), the artist catalogues what a modern computationalist might view as a list of formal state-transition operators for exploring the state-space of surrealist art. Philosophers often question whether machines can truly ground the symbols they manipulate in real world meanings (e.g. see the infamous *Chinese Room* thought experiment of Searle (1980)), but Magritte's manifesto shows how

art actively subverts this grounding to simultaneously disorient and intrigue viewers. A computer's lack of true grounding is thus less problematic in artistic creativity, and may actually serve as an advantage in crafting outputs that undermine the grounding of others.

Other challenges followed in the 1960s, with the arrival of the "beat" generation and the development of the cut-up method. Arguing that the literary arts were 50 years or more behind their visual brethren, William Burroughs and Brion Gysin invented a textual parallel to the collage, allowing writers to create new texts from old with a scissors, a paste-pot and a set of pre-existing texts to cut up (Burroughs, 1963; Lydenberg, 1987). By randomly dissecting and re-combining chunks of existing text - cut from poems, novels, newspapers or indeed anything at all - Burroughs sought to avoid the ruts and clichés that implicitly guide the writing process. As Burroughs put it, an artist "cannot will spontaneity", yet one can "introduce the unpredictable spontaneous factor with a pair of scissors". The textual cut-up is Duchamp's readymade taken to the next level with the use of a random aleatory mechanism: but rather than seek *found art* directly in the textual readymades of others, Burroughs would first create new readymades from old by cutting and splicing, before applying the same processes of selection to identify the candidates with the most intriguing emergent meanings. If what seems most natural to us eventually becomes the stuff of cliché and idiom, Burroughs and Gysin deliberately conjured up the unnatural with their use of a most unnatural generation technique.

As such, the mechanical aspects of the cut-up, which operate best when they operate without any regard to the meaning of the texts that are sliced and spliced, are easily implemented in a generative text system. But what makes the cut-up a means of creation rather than of mere generation is the selectivity of the artist. Duchamp's urinal was not art until Duchamp made it art, by selecting and framing it as such, and a random cut-up of texts is not art unless explicitly chosen as art by a creative writer who deliberately discards much more than is kept and who only keeps the combinations that work. Human creators, even those as far-seeing as Burroughs, are naturally skeptical of a machine's ability to do more than merely generate, and to critically filter its own outputs. When asked by writer William Gibson as to why he did not use a computer for his cut-ups, Burroughs replied "I dont need a computer. I have a typewriter" (Gibson, 2005).

Burroughs' cut-up method flourishes on Twitter, in the guise of automated text generators called *Twitterbots* (Veale, 2014). Most bots implement the mechanical aspects of the cut-up - the patchwork manipulation of existing texts to produce novel linguistic candidates to tweet - but omit the parts that require actual intelligence: the interpretation and filtering of candidates, to retain the best and the discard the rest. Most bots implicitly ask their human followers to provide this discernment, by favoriting or retweeting those tweets that are worthy of analysis and by ignoring the rest. Humans willingly perform this crucial part of the cut-up method for reasons that are subtle and complex. First, it allows users to share in the paternity of a creative tweet, and to serve a somewhat Duchampian role in the recognition of creative value in the found objects that are bot texts; and second, because of the relative rarity of worthy tweets, and for the uncanny feeling of seeing a system transcend its observed limitations to suddenly produce a text that is not just well-formed but surprising or even profound.

Computational Creativity, a relatively new branch of AI which sees machines not just as automated scissors or as fast typewriters, but as viable Burroughs' and Duchamps in their own right (Veale, 2012), poses the 21st century's biggest challenge to received notions about art and creativity. Yet as machines become smarter and more selective, and shoulder more of the creative burden, our machines will be challenged also: users will expect more *hits* and tolerate fewer *misses*; they will

expect the same diversity but demand more self-control. In this paper we consider the design of a next-generation Twitterbot, one that retains the ability of first-generation bots to surprise and amuse and evoke an occasional sense of the uncanny (or what (Freud, 1919) called *Das Unheimliche*), but one that also employs knowledge of words and of the world to ensure that its outputs are always meaningful, frequently apt and occasionally profound. This Twitterbot focuses on a uniquely human task, the ability to generate illuminating metaphors in which one aspect of the world is viewed through the lens of another. After conducting a survey of first-generation bots in section 2, and a review of computational approaches to metaphor in section 3, we introduce our metaphor-generating Twitterbot, @*MetaphorMagnet*, in section 4. We consider what Twitterbots of the first and next generations can tell us about the nature of general intelligence, particularly of the artificial variety and how it is received by humans, in section 5, before the outputs of @*MetaphorMagnet* are evaluated against those of a comparable first-generation bot, called @*MetaphorMinute*, in section 6. Finally, we conclude with some predictions about the future of next-generation creative Twitterbots in section 7.

2. First-Generation Twitterbots

Bots of all varieties abound on Twitter. Some bots generate novel content of their own, so that each tweet has a different form with a different potential meaning. Others employ an inventory of stock phrases, but use this inventory in ways that are mischievously apt. Some are designed to generate in a vacuum, while others are built to be responsive, eavesdropping on tweets between human users before injecting themselves into a conversation. Others are more polite, and react with apt responses to tweets directed at the bot itself. Yet almost all Twitterbots fall into a category we dub *First-Generation Bots*. These bots pair high-concept ideas with low-complexity implementations, to yield what internet artist and bot-maker Darius Kazemi calls *tiny subversions* (Kazemi, 2015). Such bots are quickly implemented, and often make a virtue of their simplicity. Yet they also shift so much of the creative burden onto their users - who must spot the occasional grains of wheat amongst large quantities of chaff and "like" or "retweet" them accordingly – that they themselves do not earn the label "creative." By exercising their own selectively in this way, it is the human followers of these bots, and not the bots themselves, that fill the role of William Burroughs in the execution of the cut-up method.

Yet the simplest bots offer the clearest insights into why humans actually follow artificial content generators on Twitter. Consider the Twitterbot @*everycolorbot*, which generates a random six-digit hex-code every hour. As each code denotes a different color in the RGB color space, @*everycolorbot* attaches a swatch of the corresponding color to each tweet. The bot's followers, which number in the tens of thousands, favorite and re-tweet its outputs not because they prefer specific RGB codes over others, but because they bring their own visual appreciation to bear on each color. They favorite or re-tweet a color swatch because of what the color says about their own unique aesthetics. Veale and Alnajjar (2015) present a variety of next-generation counterparts to @*everycolorbot*, in the form of bots that use knowledge of words and of colours (specifically, of how words denote ideas with stereotypical colorings, such as grey for *brain* or green for *alien*) to give meaningful names (such as "alienbrain" for a grayish shade of green) to the RGB hex codes generated by @*everycolorbot*. Veale and Alnajjar use crowd-sourced human evaluations to compare the names generated by their bot to those generated by human colour aficionados for the same RGB

codes on *ColorLovers.com*, and report a marked preference for the machine-generated names among human judges.

Or consider @*everyword*, which tweets the next word in its alphabetized inventory of English words every 30 minutes. (@*everyword* has since exhausted its word list, generating much media speculation (Dewey, 2014) as it neared the end of the *Z*'s.) The bot, which attracted over 100 thousand followers, tweeted words, not word meanings, yet followers brought their own context and their own meanings to bear on those tweets that occasionally (and accidently) resonated with their times. For example, the word "woman" – first tweeted on May 14, 2014 – was retweeted 243 times and *favorited* 228 times not because followers found the word itself to be new or unusual, but because the tweet coincided with the firing of the New York Times first female executive editor, in a decision that drew the ire of many for its apparent sexism. First-generation bots do not offer their own meanings, but give opportunities to their followers to impose and share meanings of their own. Timely bot tweets are conversational hooks, allowing us to show that we are in on the joke and part of the conversation.

First-generation bots explore a space of surface forms, not underlying meanings, and leave the question of meaning to their users. The mechanics of the cut-up, as exemplified by the bots @*pentametron* and @*twoheadlines*, offer as good a basis for negotiating this space as any other. @*pentametron*, by developer Ranjit Bhatnagar, crafts accidental poetry in iambic pentameter from a random pairing of tweets that match basic constraints on metre and rhyme: a tweet must have ten syllables to be chosen, and paired tweets must end with a rhyming syllable. These formal constraints lead to pairings that - in the mind of a reader at least - may give rise to fascinating emergent meanings, as in "Pathetic people are everywhere / Your web-site sucks, @*RyanAir*". @*twoheadlines*, by Darius Kazemi, employs a more traditional take on the cut-up method, and splices two current news headlines together to invent a third. The bot uses a simple but effective recombinant technique, replacing a named-entity in one headline with one from another. Though its cut-ups are sometimes jarring and blackly funny, as in *"Miss Universe attacks northeast Nigerian city; dozens killed"* (where *Boko Haram* is replaced with *Miss Universe*), most (such as *"Flights escorted to Boko Haram airport amid bomb threats"*) do not register as either witty or provocative. As long as the bot lacks knowledge of the named-entities it puts into a headline, or a means of measuring the resulting change of meaning, it can never evaluate its own cut-ups.

Bots like @*twoheadlines* generate tightly-constrained cut-ups. By inserting a single named-entity into a target headline, @*twoheadlines* aims to preserve most of its meaning by keeping most of its form. Too much random variation, coupled with a bot's inability to interpret the resulting forms, can yield senseless results. As Giora et al. (2004) demonstrate, novelty gives the most pleasure when one can make sense of it and appreciate its significance. Consider a bot named @*MetaphorMinute*, also from Darius Kazemi. This Twitterbot more or less randomly instantiates the copula-frame "An *X* is a *Y*: *P* and *Q*" with words (from Web service *Wordnik.com*) to issue a tweet that looks like a novel linguistic metaphor every two minutes. Metaphors can take many forms - unlike, say, similes - and are ultimately conceptual, rather than linguistic, in nature (Lakoff and Johnson, 1980). Though many metaphors can be expressed in a copula form *"X is a Y"* where X and Y denote incompatible ideas, it is not the case that the incompatibility of X and Y is sufficient to turn *"X is a Y"* into a metaphorical statement. Thus, though *"an astrolabe is a tapioca: lubberly yet species-specific"* may be many things to many people, it will surely be seen by very few as an apt metaphor. Nonetheless, the human mind cannot help but look for significance in any pairing of ideas, if only briefly (Milic, 1971), and the peculiarity of the words used here (e.g. the piratical

"lubberly") certainly adds to the fun. And one needn't worry about missing the meaning entirely, for like a missed bus, another tweet will arrive in a minute or two.

It is often useful to view a simple first-generation bot as the embodiment of the null hypothesis for a seemingly complex act of human creativity. *@MetaphorMinute* thus embodies the null hypothesis for creative metaphor generation, by eschewing all knowledge of humans or of the world and acting as though one can reliably generate meaningful metaphors without any representation at all of their meaning. When one is forced to reject this hypothesis, as we may well do with *@MetaphorMinute* on the basis of its often uninterpretable tweets, it becomes necessary to adopt a different hypothesis, a *knowledge* hypothesis, which claims that meaningful and intelligent metaphors cannot be crafted without world knowledge. Since this hypothesis involves a good deal more complexity, we shall empirically evaluate the resulting bot and its tweets against the null hypothesis as embodied by *@MetaphorMinute*. We begin by exploring how AI systems explicitly model metaphor as a knowledge-driven process.

3. Computational Approaches to Metaphor

Computational approaches to metaphor can most usefully be divided into four different, albeit non-disjoint categories: the *categorial*, the *corrective*, the *analogical* and the *schematic*. This heterogeneity reflects the complex, multifaceted nature of the phenomenon, for even Aristotle, who gave it its name ("phor" for *carry* and "meta" for *over*), found it necessary to offer a four-pronged analysis of metaphor in his Poetics (Hutton, 1982). The first three of Aristotle's schemata treat metaphor as a playful manipulation of taxonomic categories while the fourth sees it as a kind of proportional analogy.

3.1 Categorical Approaches

The popularity of the copula form *"X is a Y"* suggests that we naturally see metaphor as an act of *category inclusion* (Glucksberg, 1998), in which a metaphor asks us to include X in some of the same categories we place Y. Precisely which categories should be broadened to hold both Y and X is the principal focus of categorial approaches. While Aristotle was decidedly vague in his exploitation of genus and species terms, Glucksberg argues that Y serves a dual purpose in a metaphor: it denotes both its usual meaning and the category of things for which it is highly representative. Thus, in the metaphor *"my job is a jail"*, jail denotes a place of legal incarceration and the category of oppressive environments more generally. While the former adds texture to a metaphor (evoking iron bars, dark cells, brutal guards, etc.), the latter denotes the category to which job is added. Conventional metaphors exploit words with established dual meanings, like "shark" (a killer fish, or more inclusively, anything ruthless and predatory) but novel metaphors ask us to invent these inclusive categories on the fly. Way (1991) thus posits a *Dynamic Type Hierarchy* (DTH) as the centerpiece of her categorial approach, but does not describe a computational implementation of a DTH. Certainly, existing taxonomies like that of WordNet (Fellbaum, 1998) are not equal to the demands of the task. Veale and Li (2015) thus acquire a dense hierarchy of divergent categorizations from Web texts by seeking out instances of *"A_Bs such as Cs and Ds"*. This system of cross-cutting categories supports metaphor by suggesting categories that unite distant ideas such as DIVORCE and WAR (both are traumatic events and severe conflicts). Veale and Li offer this categorial approach as a free Web service, *Thesaurus Rex*, for use by third-party systems.

3.2 Corrective Approaches

Corrective approaches to metaphor tacitly assume that the language of thought is a literal language for literal thoughts. Almost any attempt to accommodate a metaphorical thought - such as that cars might *drink* gasoline - in a literal representation will thus cause a semantic anomaly that must be corrected to produce a meaning that can be literally accommodated. To this end, Wilks (1975) proposes that semantic constraints on case-frame fillers be represented not as hard restrictions but as soft preferences, while Wilks (1978) makes preferences more active, so that the most apropos literal case frame for a metaphorical utterance - the most similar frame that causes the least number of preferences to be broken - can be found. In this way, a car that drinks gasoline is accommodated in a CONSUME frame, as a car that consumes gasoline. Fass (1991) categorizes frames with an ISA hierarchy, so that the nearest and most accommodating literal frame can be found.

3.3 Analogical Approaches

Analogy, like metaphor, is a way of perceiving similarity. Each finds relational parallels in our mental organization of distant ideas, allowing us to use our knowledge of one idea to reason about another. Aristotle categorized proportional analogies of the form *A:B::X:Y* as a type of metaphor, while computational analogists such as Gentner, Falkenhainer, and Skorstad (1989) view metaphor as a special application of analogical principles. But rather than seek a single *A:B::X:Y* proportion, analogists seek to build coherent systems of proportions, to construct an isomorphic mapping from the graph representation of a source domain to the graph representation of a target (Falkenhainer, Forbus, and Gentner, 1989; Veale and Keane, 1997). Thus, an analogical reading of *"my car drinks gasoline"* might map CAR onto DRINKER, GAS onto ALCOHOL and the side-effects of heavy drinking (unreliability, constant craving, wastefulness) on to key aspects of driving (reliability, mileage, fuel efficiency). Such examples demonstrate both the power and the price of computational models of analogy: they achieve deep results only when they have rich representations to work with. To acquire the necessary consensus knowledge of everyday ideas, Veale and Li (2011) harvest widespread beliefs, such as that religions ban alcohol and that businesses pay taxes, from *why do* questions posed to Web search engines. While Google does not expose its query logs, it does expose the most common *why do* questions in the form of helpful query completions. Completions for *Why do Xs* expose the shared beliefs that users suppose everyone else to think about X, making these beliefs ideal for figurative processing. Veale and Li provide a Web-service called *Metaphor Eyes* for generating and interpreting metaphors with this knowledge; this service may be freely used by third-party systems.

3.4 Schematic Approaches

A jarringly mixed metaphor can remind us that metaphors work with each other more often than against each other. Yet many thematically-similar metaphors for familiar ideas - such as LIFE, LOVE, ANGER and DEATH - may not actually be distinct metaphors at all, but the products of a common deep structure that manifests itself at the word level in diverse but inter-operable ways. Lakoff and Johnson (1980) thus argue that metaphor is fundamentally a conceptual phenomenon in which linguistic metaphors, such as "hitting the rocks" and "crashing and burning", are the surface realizations of an underlying metaphor schema such as LIFE IS A JOURNEY. This *Conceptual Metaphor Theory* (CMT) suggests that a finite number of productive metaphor schemas

are responsible for an unbounded number of linguistic metaphors, from the old and conventional to the new and the poetic. Computational approaches to CMT, as offered by Carbonell (1981), Martin (1990) and Barnden (2008), are most effective when dealing with the conventional metaphor schemas that are ubiquitous in everyday language, as these may take many forms and underpin a great many habitual thought patterns. But as we show next, the schematic approach can also be productively used for the generation of novel metaphors.

4. Metaphor Generation in *@MetaphorMagnet*

One cannot speak knowledgeably without knowledge, and metaphor is nothing if not a knowledgeable way to put ideas into words. The metaphor-tweeting bot *@MetaphorMagnet* thus exploits a variety of knowledge sources. For categorial knowledge, it draws upon the public Web service *Thesaurus Rex* (Veale and Li, 2015). This service provides diverse, fine-grained categorizations of any topic on demand, noting that e.g. NICOTINE and COFFEE are each, to differing degrees, toxic, addictive and psychoactive substances. It also uses another Web service, *Metaphor Eyes* (Veale and Li, 2011), to obtain the relational knowledge it needs to make analogies. Thus, for instance, it combines knowledge from both of these services to generate the following provocative tweet:

> Would you rather be:
> 1. A psycho producing an obnoxious rant?
> 2. A wordsmith producing a charming metaphor?
> *#Psycho = #Wordsmith*

Here it is *Thesaurus Rex* that indicates that rants are often considered *obnoxious* while metaphors are often *charming*, while *Metaphor Eyes* is the source of the triples $psycho-produce-rant$ and $wordsmith-produce-metaphor$. While obnoxious rants and charming metaphors are each uttered by very different kinds of speaker (the positive-affect "wordsmith" versus the negative-affect "psycho"), the central relationship $S-produce-U$ is the same for each. *@MetaphorMagnet* constructs each figurative tweet upon an affective contrast such as *obnoxious:charming*, using the knowledge at its disposal to motivate and elaborate this contrast before employing one of a wide range of rhetorical formulations to package the conceit in a pithy and provocative fashion. With the right prompt, many such contrasts might even rise to the level of an ironic observation on life. *@MetaphorMagnet* nudges its followers to see the potential irony with an oft-used Twitter practice: by appending the hashtag *#irony*.

> *#Irony*: When savory steaks sizzle like the most unsavory scandals.
> *#Savory = #Unsavory #Steak = #Scandal*

A further linguistic marker of irony is the "scare" quote, as in:

> *#Irony*: When some hosts arrange "entertaining" parties the way
> presenters arrange boring lectures.
> *#Host = #Presenter #Party = ł#Lecture*

@MetaphorMagnet's musings do not reflect real experience, only the formal possibilities of such. It has never been bored at a party, it has never eaten a sizzling steak, and it has never enjoyed

a juicy scandal. Nonetheless, this recognition of formal possibility is enough to spur its readers to lend their own experience to its tweets. The key lies in using knowledge to meet its readers half-way. @*MetaphorMagnet* thus brings a third source of knowledge to the table: knowledge of famous figures, real and fictional (Veale, 2015). With a bespoke knowledge-base that uses more than 30,000 triples to describe more than 800 figures, from *Plato* to *Darth Vader*, @*MetaphorMagnet* creates bizarre but apt pop-cultural analogies such as the following:

> #NevilleChamberlain is the #FredoCorleone of #WorldWarII:
> weak and gullible, yet well-meaning too.

The central contrast here, between the real and the fictional, is elaborated in a follow-up tweet from @*MetaphorMagnet*:

> If #NevilleChamberlain is just like #FredoCorleone, weak and gullible,
> then who in #TheGodfather is #WinstonChurchill most like?
> #WorldWarII

While the system cannot know a figure like *Fredo Corleone* to the same degree as one who has watched and enjoyed (or even understood) *The Godfather*, it uses what it does know effectively and sparingly. Though it can answer its own question here (an apt answer might be *Michael Corleone*) it prefers to hold back, so as to draw the reader further into its tweets.

@*MetaphorMagnet* also employs a schematic approach to metaphor generation, though rather than code any schemas explicitly, it opportunistically harvests a wide range of likely schemas from the Google Web n-grams (Brants and Franz, 2006). As the Web is replete with articles about CMT, most of CMT's schemas are to be found in the Google 3-grams, where e.g. "ARGUMENT IS WAR" has a frequency of 275 and "TIME IS MONEY" has a frequency of 14179, or in the 4-grams, where "LIFE IS A JOURNEY" has a frequency of 4130. Web n-grams are also a rich source of folk schemas, those copula metaphors that pervade a culture because of a popular song, movie or book (e.g. "LOVE IS A BATTLEFIELD", freq=7290, or "LIFE IS A CABARET", freq=2184). Whatever the provenance of a likely schema, @*MetaphorMagnet* uses it if it can. To lend new resonance to these linguistic readymades, @*MetaphorMagnet* pairs them in contrastive juxtapositions, for if schemas offer alternate ways of seeing, a clash of schemas can alert us to a conflict of world views, as in a tweet that finds discord in two sharply contrastive views of the law:

> To some legislators, law is a placid mirror.
> To others, it is a devastating tyranny.
> *#Law = #Mirror #Law=#Tyranny*

As the hashtags reveal, some say "LAW IS A MIRROR" (4-gram freq=64) and others say "LAW IS TYRANNY" (3-gram freq=67). @*MetaphorMagnet* does not go so far as to reason that laws hold up a mirror to society, or serve as the tools of tyrants; that is for a reader to conclude. Rather, the bot simply meets its readers half-way by packing two contrasting views on the law, the *placid* versus *devastating*, into a coherent tweet. These associations come from *Thesaurus Rex*, while the fact that legislators write laws comes from *Metaphor Eyes*. The clash of world views gives the metaphor its desirable tension, but this tension would be greater still if it were made social, by imagining the kinds of users that might espouse those views. Twitter users often choose their handles to reflect

their views of the world, allowing a deft bot to invent an apt name for an imaginary user from the conceptual metaphor they appear to live by. The following tweet from *@MetaphorMagnet* turns a clash of metaphors into a fictional debate, Twitter-style:

> .@*war_poet* says history is a straight line
> .@*war_prisoner* says it is a coiled chain
> *#History = #Line #History = #Chain*

While the handle *@war_poet* names an existing Twitter user, it is reinvented here by *@MetaphorMagnet* from the 2-gram "war poet" (freq=2688) to personify the view that "history is a line" (4-gram freq=89), using *Metaphor Eyes* to provide the facts that poets write lines and that histories record wars. It is a name that imbues the metaphor HISTORY IS A STRAIGHT LINE with added resonance, though more resonance still is found in the handle *@war_prisoner*, which lends a sinister hue to the idea of history as a cycle that one cannot escape. While *@MetaphorMagnet* is not yet capable of this level of critical analysis, it is responsible for creating the kind of metaphors that support just this level of analysis by interested readers.

5. Crashing the Party: Verbal Acumen and General Intelligence

What can Twitterbots tell us about general intelligence, in humans or in artificial systems? At first glance, it would seem that first-generation bots have little to tell us about human intelligence or understanding, since such systems are skillfully designed to push questions of interpretation and meaning onto their human followers. Yet, while shuffling words so as to create the impression of meaning where none is possessed may seem like a distinctly mechanical hack, it is a strategy that is not uncommonly employed by humans too. For instance, *cocktail party syndrome* (also known as *chatter-box syndrome*) is a medical condition observed in children with arrested hydrocephalus whose avid sociability, loquaciousness and apparent verbal acumen conceal a diminished general intelligence, allowing them to speak with confidence and apparent knowledge about topics they know nothing about, using words whose actual meaning is lost on them. As noted in (Schwartz, 1974, p.466), "the child uses automatic phrases and clichés; at times he even quotes directly from television commercials or slang he has heard others use. He uses words from other contexts that almost but not quite fit his conversation." One wonders what Searle (1980) might make of such a child and his ungrounded use of symbols to appear intelligent. For to substitute an AI analogy for a medical metaphor, the affected child acts as much like a first-generation bot as a cocktail-party bore. Indeed, though the condition has a physiological basis in children (*hydrocephalus*), its name suggests that such behaviour is often observed in adults too, who have no such medical diagnosis to explain their engaging but unintelligent use of words in convivial social settings. Twitterbots (or chatterbots generally), much like human chatterboxes, exploit our very human willingness to attribute meaning and intent to linguistic forms that obey the conventions of language, even if one has to work so hard to unearth those meanings that one is effectively inventing them for oneself.

As Reddy (1979) explains, the *Conduit metaphor* of human communication – that folk model of language that sees our words and phrases as containers of meaning that shuttle back and forth between speakers - is much more a convenient fiction than it is a cognitive reality. Our words are not containers to be loaded with meaning at one end of a channel so as to be unloaded by a listener at the other; rather, we use words as structured prompts to spur the creation of meanings in the minds of an audience. Good speakers use these prompts well, to spur a listener to recreate

something approximating the meaning or the feeling the speaker has in mind, if not more besides. But not everyone that speaks well can communicate well, for humans can create the impression of being adept with words while failing to communicate one's thoughts or ideas well. Consider this description in McKeganey (1983) of a child, *Linda*, who exhibits all the signs of cocktail-party syndrome: "Linda lives in language and loves to talk and listen to it. She does not however always grasp the meaning and is inclined to indulge in the sound and play of words." Though Linda seems a bright and sociable child, her attitude to language is not so very different (though orders of magnitude more complex, it must be said) to a Twitterbot such as *@Pentametron*. Such bots, like Linda, seem to reside "inside language" rather than "in the world". Linda is a child who uses language to fit in rather than to communicate, leading a doctor (cited by McKeganey) to note of her that she "Talks like a grown up Yankee. Incredibly charming. Incredibly vulnerable. Adult language [but] infantile frustration threshold." McKeganey adds that Linda experiences (and causes) frustrations because of her tendency to act like an adult while having the mind of a child.

If most first-generation Twitterbots are technological children that suffer from the machine equivalent of cocktail-party syndrome, the goal of next-generation bots is to give our machines something meaningful to talk about, and the wherewithal to put these meanings into words so they can be good communicators as well as good speakers. But if it is easier to see the symptoms of cocktail-party syndrome in a child like Linda than it is in an adult at an actual cocktail party, it is not always so easy to see the difference between a first- and a next-generation Twitterbot, that is, between a bot that "lives in language" and one that resides in the world, or at least in a logical model thereof. McKeganey argues that it is context that shows up Linda's inappropriate use of adult language and ideas, and it is an insensitivity to context that leads others to see her use of language as lacking in real meaning. When tweets are shorn of context, we can expect the outputs of first-generation bots to appear every bit as intelligent - if not more so - than those of their next-generation brethren, no matter how much knowledge and general AI we put into the latter. In the next section we conduct a side-by-side evaluation of one instance of each generation of bot, to determine the willingness of human raters to judge the value of a linguistic artifact on its form as well as on its content. As we shall show, a clever use of form can seduce the reader into seeing meaning where none was intended. To get human judges to look past the allure of form and judge a text on its meaning, we shall have to make them work to recover this meaning for themselves.

6. Bot Vs. Bot: A Comparative Evaluation

@MetaphorMagnet and *@MetaphorMinute* embody very different hypotheses about metaphor creation. The former embodies the belief that knowledge truly matters; a system for generating novel metaphors must represent enough about words and its ideas to use either of them in intelligent and interesting ways. The latter embodies the corresponding null hypothesis, the belief that metaphor is a matter of form, not knowledge, a form that readers will fill with their own personal meanings. As neither bot has yet garnered enough responses for robust statistical analysis of their rates of favoriting and retweeting, we thus use the crowdsourcing platform *CrowdFlower.com* to elicit human ratings for the outputs of each. For our first experiment, we randomly sampled 60 tweets from each bot, and solicited 10 user ratings (along 3 dimensions) for each on CrowdFlower. Raters were paid a small sum per rating. They were not told that the tweets were produced by bots, but simply that each was a metaphor observed on Twitter. The three dimensions that were rated are *Comprehensibility*, *Novelty* and *Re-Tweetability*. For each tweet, raters were asked to

provide a rating from 1 to 4, where $1 = VeryLow$, $2 = MediumLow$, $3 = MediumHigh$ and $4 = VeryHigh$ quality.

Table 1 presents the distribution of mean ratings for the *Comprehensibility* of each Twitterbot's tweets. Over 51% of *@MetaphorMagnet*'s tweets are deemed to have very high comprehensibility, and only 25% are rated as being hard or somewhat hard to understand. *@MetaphorMinute*'s outputs are rated as much harder to comprehend, yet over 50% of its tweets are still deemed to have medium- to very- high comprehensibility. Seduced by form, human raters are often willing to see meaning where no specific meaning was ever intended.

Comprehensibility	*@MetaphorMagnet*	*@MetaphorMinute*
Very Low	11.6%	23.9%
Medium Low	13.2%	22.2%
Medium High	23.7%	22.4%
Very High	51.5%	31.6%

Table 1: Comparative evaluation of *Comprehensibility* ratings

Table 2 shows the distribution of mean ratings for the perceived *Novelty* of each bot's tweets. *@MetaphorMinute* outdoes*@MetaphorMagnet* on ratings of very high novelty, perhaps because it presents raters with unnatural combinations of words that are so very rarely seen in human-crafted texts. Since raters can also confidently assign a meaning to 31.6% of *@MetaphorMinute*'s random confections, we may infer that humans follow the Twitterbot because a significant portion of its outputs can be read as being very novel yet somewhat meaningful.

Novelty	*@MetaphorMagnet*	*@MetaphorMinute*
Very Low	11.9%	9.5%
Medium Low	17.3%	12.4%
Medium High	21%	14.9%
Very High	49.8%	63.2%

Table 2: Comparative evaluation of *Novelty* ratings

However, Table 3 shows the distribution of mean ratings for the expected *Re-Tweetability* of each bot's outputs. Simply, annotators were asked for the likelihood that they would re-tweet the given metaphor. Raters are not generous with their *Very-High* ratings for either bot, but most Twitter users can be just as sparing in their use of re-tweeting. Indeed, since raters were only asked to speculate about a metaphor's re-tweet value, we can expect the mean ratings of Table 3 to be somewhat higher than actual rates of re-tweeting on Twitter. Nonetheless, *@MetaphorMagnet*'s metaphors are rated, on average, to have significantly higher re-tweet value (where 2 in 5 are deemed to have medium-high to very-high retweetability) than those of *@MetaphorMinute* (where just 1 in 4 is deemed to have medium-high to very-high retweet value).

A second experiment was conducted to evaluate the perceived aptness of each bot's outputs. The same number of tweets was newly sampled from each bot, and the same number of ratings was solicited from human raters. However, to spur the human judges into actually looking for the meaning of each tweet, a cloze test format was now used, so that a pair of pivotal qualities

Re-Tweetability	*@MetaphorMagnet*	*@MetaphorMinute*
Very Low	15.5%	41%
Medium Low	41.9%	34.1%
Medium High	27.4%	15%
Very High	15.3%	9.9%

Table 3: Comparative evaluation of *Re-Tweetability* ratings

was blanked out from each text presented to the judges. For example, the focal words *spoiled* and *whipped* (shown underlined below) were replaced with blanks in:

> So I'm not the most spoiled pooch in the kennel.
> More like the most whipped hound in the pack.
> *#Kennel = #Pack*

@MetaphorMinute's tweets were similarly templatized, so e.g. the qualities *anchored* and *excusatory* (shown underlined below) were blanked out in the following tweet:

> a provisor is a symphysis: anchored and excusatory

For each tweet from each bot, judges were shown five pairs of qualities to choose from so as to re-fill the blanks with apt content. These five pairs were ordered randomly, and included the original pair of qualities (e.g. spoiled and whipped) and four pairs of distractors, taken from other outputs of the same bot. Raters chose the right (original) pair for *@MetaphorMagnet*'s tweets 60% of the time. In contrast, raters choose the correct pair for*@MetaphorMinute* no more often than random chance would predict, or just 15% of the time. Table 4 shows the distribution of tweet *Aptness* ratings for each bot, where a metaphor is said to have *Very-Low* aptness if the original pairing is chosen 25% of the time or less by raters, *Medium-Low* if chosen no more than 50% of the time, *Medium-High* if chosen 51-75% of the time, and *Very-High* if chosen over 75% of the time.

Aptness	*@MetaphorMagnet*	*@MetaphorMinute*
Very Low	0%	84%
Medium Low	22%	16%
Medium High	58%	0%
Very High	20%	0%

Table 4: Comparative evaluation of *Aptness* ratings

A metaphor demands both novelty and aptness, for it is aptness that allows novelty to serve a descriptive purpose, and it is novelty that causes this aptness to seem fresh to a reader. Novelty alone is easily achieved with aleatory methods (e.g., see Chamberlain and Etter (1983)'s *Racter* system), but aptness requires a mutual understanding of words and the world so that readers can be drawn into the shared world of the metaphor. Though there is an observable *placebo effect* at work in first-generation bots, where readers are seduced into perceiving meaningfulness (though not necessarily meaning) on the basis of form alone, a bot can achieve only so much with a knowledge-free approach that relies so heavily on the kindness of its followers.

7. Concluding Remarks: The Kindness of Strangers

In its first year of activity, *@MetaphorMagnet* has tweeted approx. 9000 figurative tweets, at a rate of one per hour. As a public Twitterbot, all of *@MetaphorMagnet*'s outputs - its *hits* and its *misses* - are open to very public scrutiny. So too are the responses of other, human users. The following are especially revealing about our complex expectations of AI bots:

> *@rkatz00* what i am most interested in re:*@MetaphorMagnet*
> is all the generating of wild equivalences... metaphor=jawbreaker eyesocket=braintoilet

> luke *@luke_barnes* Oct 17 *@MetaphorMagnet*
> Guys a COMPUTER PROGRAM wrote this

The latter was a response to the following figurative tweet from *@MetaphorMagnet*:

> Beauty inspires the faith that promotes courage.
> Uncertainty promotes the hope that promotes courage.
> Take your pick.
> *#Beauty = #Uncertainty?*

We value bots for their ability to speak to human concerns like hope and courage without ever pretending to be human. We want bots to generate coherent meanings, but for those meanings to come from a place of emotional detachment, to provide uncanny fodder for further human consideration. So the following tweet, which seems childlike in its curiosity, also mixes cold logic with an unnerving lack of empathy:

> What if a world with more bigots is a world with less ghettoes
> since bigots attack the minorities that live in ghettoes?
> *#Bigot*

A human reader may take this musing to be ironic, or not. Bots, like artists, should not aim to be right, just interesting. As Susan Sontag remarked in her essay *Against Interpretation* (Sontag, 1966, p.8), "Real art has the capacity to make us nervous."

Max Ernst defined visual collage - a surrealist forerunner to Gysin and Burroughs' cut-up method – as "the systematic exploitation of the fortuitous or engineered encounter of two or more intrinsically incompatible realities." Many first-generation bots build inter-textual collages, yet place greater reliance on the fortuitous aspects of a linguistic encounter than on systematic knowledge engineering. Indeed, because first-generation bots lack a semantic representation of the domains in which they work, they lack the intelligence to realize that two realities are intrinsically incompatible. This *sine qua non* of a thought-provoking collage is thus largely left to chance. *@MetaphorMagnet* shows how next-generation Twitterbots might reliably exploit bottom-up, data-driven, aleatory methods within a top-down knowledge-based approach to intelligent content generation. Artists make their own luck, by recognizing a fortuitous encounter when they see one, or indeed, when they generate one, and so our bots must also possess the knowledge to see the potential for real meaning and palpable tension in the texts that they generate. It is this grounding in world knowledge, and the ability to give the inferences derived from this knowledge a satisfying linguistic form, that separates speakers of the cocktail-party variety - whether human or artificial - from communicators with genuine intelligence and wit.

References

Barnden, J. 2008. Metaphor and artificial intelligence: Why they matter to each other. In Gibbs, R. W., ed., *The Cambridge Handbook of Metaphor and Thought*. Cambridge, UK: Cambridge University Press. 311–338.

Brants, T., and Franz, A. 2006. *Web IT 5-gram database, Version 1*. The Linguistic Data Consortium.

Burroughs, W. S. 1963. The Cut-Up Method. In Jones, L., ed., *The Moderns: An Anthology of New Writing in America*. New York, NY: Corinth Books.

Carbonell, J. G. 1981. Metaphor: An inescapable phenomenon in natural language comprehension. Technical Report 2404, Carnegie Mellon Computer Science Department, Pittsburgh, PA.

Chamberlain, W., and Etter, T. 1983. *The Policeman's Beard is Half-Constructed: Computer Prose and Poetry*. London, UK: Warner Books.

Dewey, C. 2014. What happens when @everyword ends? *The Wall Street Journal (Intersect column)* May 23.

Duchamp, M. 1917. The Richard Mutt Case. *Blind Man* 2:4–5.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence* 41:1–63.

Fass, D. 1991. Met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics* 17(1):49–90.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Freud, S. 1919. Das Unheimliche. In *Collected Papers*, volume XII. G.W. 229–268.

Gentner, D.; Falkenhainer, B.; and Skorstad, J. 1989. Metaphor: The Good, The Bad and the Ugly. In Wilks, Y., ed., *Theoretical Issues in Natural Language Processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, W. 2005. God's Little Toys. *Wired Magazine* 13.07.

Giora, R.; Fein, O.; Kronrod, A.; Elnatan, I.; Shuval, N.; and Zur, A. 2004. Weapons of Mass Distraction: Optimal Innovation and Pleasure Ratings. *Metaphor and Symbol* 19(2):115–141.

Glucksberg, S. 1998. Understanding metaphors. *Current Directions in Psychological Science* 7:39–43.

Hughes, R. 1991. *The Shock of the New: Art and the century of change*. London, UK: Thames and Hudson.

Hutton, J. 1982. *Aristotle's Poetics*. New York, NY: Norton.

Kazemi, D. 2015. TinySubversions.com. Web-site.

Kuenzli, R. E., and Naumann, F. M. 1989. *Marcel Duchamp: artist of the century*. Cambridge, MA: MIT Press.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago, Illinois: University of Chicago Press.

Lydenberg, R. 1987. *Word Cultures: Radical theory and practice in William S. Burroughs' fictio*. Chicago, Illinois: University of Illinois Press.

Magritte, R. 1929. Les Mots et les Images. *La Révolution surréaliste* (12):32–33.

Martin, J. H. 1990. *A Computational Model of Metaphor Interpretation*. San Diego, CA: Academic Press.

McKeganey, N. P. 1983. Cocktail party syndrome. *Sociology of Health and Illness* 5(1):95–103.

Milic, L. T. 1971. The possible usefulness of computer poetry. In Wisbey, R., ed., *The Computer in Literary and Linguistic Research*, Publications of the Literary and Linguistic Computing Centre. Cambridge, UK: University of Cambridge.

Reddy, M. J. 1979. The conduit metaphor: A case of frame conflict in our language about language. In Ortony, A., ed., *Metaphor and Thought*. Cambridge, UK: Cambridge University Press. 284–310.

Schwartz, E. R. 1974. Characteristics of Speech and Language Developments in the child with Myelomeningocele and Hydrocephalus. *Journal of Speech and Hearing Disorders* 39(4).

Searle, J. R. 1980. Mind, Brains and Programs. *Behavioral and Brain Sciences* 3:417–457.

Sontag, S. 1966. *Against Interpretation and Other Essays*. New York, NY: Farrar, Straus and Giroux.

Taylor, M. R. 2009. *Marcel Duchamp: Étant donnés (Philadelphia Museum of Art)*. New Haven, Connecticut: Yale University Press.

Veale, T., and Alnajjar, K. 2015. Unweaving the Lexical Rainbow: Grounding Linguistic Creativity in Perceptual Semantics. In Ventura, D., ed., *Proceedings of ICCC-2015, the 6th International Conference on Computational Creativity*. Park City, UT: Association for Computational Creativity.

Veale, T., and Keane, M. T. 1997. The Competence of Sub-Optimal Structure Mapping on 'Hard' Analogies. In *Proceedings of IJCAI'97, the 15th International Joint Conference on Artificial Intelligence. Nagoya, Japan*. San Mateo, CA: Morgan Kaufmann.

Veale, T., and Li, G. 2011. Creative Introspection and Knowledge Acquisition. In Burgard, W., and Roth, D., eds., *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. San Francisco, CA: AAAI Press.

Veale, T., and Li, G. 2015. Distributed Divergent Creativity: Computational Creative Agents at Web Scale. *Cognitive Computation* (May):1–12.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. London, UK: Bloomsbury.

Veale, T. 2014. Running With Scissors: Cut-Ups, Boundary Friction and Creative Reuse. In Lamontagne, L., and Plaza, E., eds., *Proceedings of ICCBR-2014, the 22nd International Conference on Case-Based Reasoning*.

Veale, T. 2015. Game of Tropes: Exploring the Placebo Effect in Computational Creativity. In Ventura, D., ed., *Proceedings of ICCC-2015, the 6th International Conference on Computational Creativity*. Park City, UT: Association for Computational Creativity.

Way, E. C. 1991. *Knowledge Representation and Metaphor*. Studies in Cognitive systems. Amsterdam, the Netherlands: Kluwer Academic.

Wilks, Y. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence* 6(1):53–74.

Wilks, Y. 1978. Making Preferences More Active. *Artificial Intelligence* 11(3):197–223.