A Rap on the Knuckles and a Twist in the Tale

From Tweeting Affective Metaphors to Generating Stories with a Moral

Tony Veale

School of Computer Science, University College Dublin
tony.veale@UCD.ie @MetaphorMagnet @MetaphorMirror @BestOfBotWorlds

Abstract

Rules offer a convenient means of limiting the operational scope of our AI programs so as to not transgress predictable moral boundaries. Yet the imposition of an operational morality based on mere rules will not turn our machines into moral agents, just the unthinking tools of moral designers. If we are to imbue our machines with a profound functional morality, we must first gift them with a moral imagination, for empathic morality – where one agent treats another as it would want to be treated itself – requires an ability to project oneself into the realms of the counterfactual. In this paper we thus explore the role of the moral imagination in generating new and inspiring stories. The creation of novel tales with a built-in moral requires that an artificial system possess the ability to guess at the morality of characters and their actions in novel settings and events. Our moralizing tale-spinner – which generates Aesop-style tales about human-like animals with identifiable human qualities – also faces another challenge: it must render these tales as micro-texts that can be distributed as tweets. As we shall also use metaphor to lend elasticity to our moral conceptions, these short stories, rich in animal metaphors, will comprise part of the daily output of the *@MetaphorMagnet* Twitterbot.

Navigating the Moral Maze with AI

Morality is a complex human construct that can be viewed through the prism of various AI paradigms, each conducive to a different application or goal (Wallach & Allen, 2008). Legalistic morality is most amenable to structure-mapping models of analogy (Hunter, 2008), insofar as an analogical mapping from a new dilemma to an older precedent allows an agent to reach a moral conclusion that is consistent with one's earlier judgments. Models of conceptual blending (Fauconnier & Turner, 1998; Veale & O'Donoghue, 2000) support empathic morality, by enabling us to figuratively project ourselves into the shoes of another cognitive agent, and thereby imagine how we ourselves would react if faced with a similar dilemma. Even good-old-fashioned AI (or GOFAI) has a role to play in artificial moral reasoning, as we often imagine morality to be a guiding influence in our navigation of the space of actions with a moral dimension – the so-called *moral maze*. And then there are rules. Western culture – from Moses to Freud – offers us many reasons to see morality as a matter of rules and taboos. For instance, the ten commandments of the Judeo-Christian tradition are viewed by many as the bedrock of a moral life: ten simple rules of the "*thou shalt*" variety that set severe limits on what we must or must not do in our daily lives. This view of morality as a regulator for our human desires has obvious parallels with the modern Freudian conception of the *superego* and the *id* (Freud, 1933:107): the Freudian id is pure instinct that "knows no judgments of value: no good and evil, no morality." It must be kept in check by the *superego*, a censorious filter that one acquires via moral instruction and which acts as a moral policeman.

These AI perspectives share hidden similarities, as each is founded on an elastic category system made flexible by metaphor, analogy and blending. Consider moral rules: though written as clear-cut proscriptions of one's behavior, such rules are a good deal spongier than they seem. The ten commandments of the Bible pivot upon ideas such as *wife*, *neighbor*, *father* and *mother* in ways that are all designed to stretch, or else they could never accommodate a growing society. *Neighbor*, for instance, is a highly contextualized notion, and surely means more than the literal "person next door." Likewise, the *wife* of the ten commandments now means sexual partner (of either gender), while the goods that one might covet now include intangible assets such as status, rights, and even ideas. The notions of *father* and *mother* may be generalized to denote any senior figure deserving of respect, while the *theft* of "thou shalt not steal" now also embraces notions of intellectual property. And though immoral acts such as slavery are not explicitly prohibited, one can figuratively view slavery as the theft of one's freedom, and repression as the theft of one's voice.

Rules are of little use without illustrative use cases. The

Bible and other fonts of moral instruction are thus replete with stories that show moral rules – or violations thereof – in action, in a wide diversity of social contexts. Children's tales, from Aesop to the brothers Grimm, are likewise rich in novel scenarios that foster a clear linkage between moral reasoning and the creative mind. For it takes imagination to apply moral rules, and experience to know how far one can stretch the underlying ideas via metaphor. In this paper we set out to construct a spinner of moral "truths", whether as pithy metaphors and blends or as long-form stories with a moralistic aim. The outputs of this automated moralizer are concisely packaged as tweets to be shared by a Twitterbot named *@MetaphorMagnet* (Veale, 2014). The pairing is an synergistic one: *@MetaphorMagnet* provides the figurative flexibility a cognitive agent needs to construct provocative *bisociations* of familiar ideas in the vein of Koestler(1964).

Figurative Equivalence & Moral Metaphors

Though it is always possible to talk of absolute morality in any situation, it is often difficult for any two of us to agree on what these moral absolutes should be (e.g. consider the polarized positions of the abortion debate). In the absence of moral certainty, we find it easier to draw equivalences between situations that we see as morally "similar". Such equivalences can, ironically, lend our assertions a stronger sense of moral certainty (e.g. "*meat is murder*", "*marriage is slavery*") that disguises the very absence of certainty. In fact, one needs no moral values at all to draw equivalences between situations that are freighted with moral possibility: so machines can, for instance, draw analogies between two human actions that it judges – via sentiment analysis – to have non-neutral valence, such as *marriage* or *slavery*. In this way, a machine lacking a sense of morality can assert equivalences that provoke a moral judgment from humans. This is the primary goal of *@MetaphorMagnet*: to use figurative devices to frame the pairings of ideas that are most likely to stir the imagination, as in the following examples:

> *Justice applies the laws that reduce freedoms. Racism leads to the slavery that competes with freedom. Take your pick*. #Justice=#Racism?
>
> *.@sex_lover says marriage is a crusading vow*
> *.@sex_traitor says it is a cynical betrayal*
>
> *Spouses embrace marriage. Prostitutes profit from the sex that nurtures marriages. Who is better?*
>
> *I used to think of myself as a graceful bride that embraced marriage. Now I see myself as a clumsy buffoon that suffered from confusion.*
>
> *So I'm not the most charming flower in the meadow. More like the most charming rat in the sewer.*

@MetaphorMagnet constructs its moral equivalences using a variety of value-free strategies. Veale & Valitutti (2014) present a strategy of causal equivalence, whereby a moral equivalence is drawn between two ideas of opposing affect that can be argued to produce the same affective outcome. Thus, in the first example above, as Justice and Slavery each limit human freedoms, *@MetaphorMagnet* suggests that they might be seen by some as morally equivalent (as in "*slavish obedience to the law*"). Another strategy, described in Veale (2015), pits two conceptual metaphors for the same idea against each other. Thus, the positive view of marriage as a vow (of a zealous crusader) conflicts with a negative view of marriage as a betrayal (by a cynic). In each case, these individual metaphors are extracted from the Google 4-grams (Brants & Franz, 2006), while the bot invents Twitter handles for the champions of these views by mining apt noun-phrases from the Google 2-grams. Of course, *@MetaphorMagnet* lacks the moral values to take a side in either of these competing assertions and metaphors. Metaphor is not a question of truth but of perspective, and the bot aims to provoke human debate as a neutral outsider.

Sentiment & Morality: Naggers With Attitude

Insofar as morality is often perceived as matter of attitude, it can be profoundly shaped by the use of oratorical style. For a morality-free agent (whether human or artificial) can successfully communicate a strong moral tone by aping the recognizable style of a moralizing speaker. Twitter proves itself a fertile medium for parodies of the tics and tropes of famous authors or thinkers, as evidenced by tweets tagged with #JamesEllroyStarWars and #ThingsJesusNeverSaid. The latter hashtag is sufficiently ambiguous to be attached to tweets that either attack Christian values or that defend them from attack by liberals. Though it helps to possess a knowledge of Christian morality, it suffices to know the dominant tropes of Christian proselytization. For example, *@MetaphorMagnet* generates tweets such as the following:

> *"Verily, it is better to be a pastor working in an ornate church than a pauper working in a spartan shack." #ThingsJesusNeverSaid*
>
> *Blessed are the lowly addicts that are cured in rehab, for they shall inherit the throne and become honored princes*. *#ThingsJesusNeverSaid*

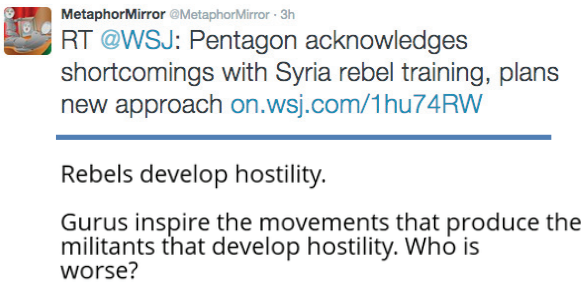
Some moralistic tones are more identifiable than others, especially if syntax is tortured in an idiosyncratic manner. Consider *@MetaphorMagnet*'s playful use of syntax here:

> *Weak is the force in soft lovers if love they fall out of*
> *Perform painful acts, they will*
> *Harsh torturers will they be*
> #ThingsYodaNeverSaid

In each case, *@MetaphorMagnet* uses its knowledge of lexical affect as a substitute for moral insight: so ideas that

are lexicalized as negative-affect words (e.g. *lowly*, *addict*, *perfidy*, *crooked*, *liar*) are deemed to be morally dubious, while ideas lexicalized as positive-affect words, such as *cured*, *throne*, *honored* and *prince*, are assumed to be morally virtuous. The bot then constructs a mini-story that takes a given character (e.g. a *liar*, an *addict*) from a negative to a positive state, and frames this change as a virtuous – and thus morally inspiring – transformation.

@MetaphorMagnet's affective metaphors can, if paired to breaking news, offer a moralizing view on current events. A companion bot, named *@MetaphorMirror*, thus pairs apt metaphors to news headlines automatically, via a similarity measure based on *Latent Semantic Analysis*. When the pairing of metaphor to headline is a strong one, the result can yield a moral judgment on the news, as in:



Note the linguistic framing here: by asking *Who is worse?* the system aims to elicit a moral judgment from its human followers even though it is incapable of making one itself.

From the Ideal Role-Fillers to Moral Values

Lexical affect offers a crude proxy for moral value, but it lacks nuance and perspective, for different participants to an action may hold conflicting views on its morality. Thus, for instance, a criminal is likely to construe the morality of his actions very differently from his victims. To judge the morality of any action, an agent requires an understanding of how participants *may* contribute to its execution and an opinion as to how they *should* contribute. The former is part of an agent's semantic representation of an action, the latter part of its value system. A cognitive agent must use both of these systems together to navigate the moral maze.

We imbue *@MetaphorMagnet* with a sense of semantic possibility by giving it a network of actions and roles in the vein of *FrameNet* (see Baker et al., 1998). So this network codifies the expectation that criminals commit their crimes *against victims*, terrorists perpetrate outrages *for mullahs*, laborers do work *for bosses,* conmen pull tricks *on dupes,* surgeons operate *on patients*, and so on. One may execute the responsibilities of one's role very well or very badly, so a dim-witted accomplice or a loose-lipped conspirator may deserve the opprobrium of a protagonist regardless of the morality of the action in which they are jointly engaged.

For over 300 roles, from *victim* to *worshipper* via *rival*, *follower*, *sympathizer*, *boss, patient, dupe* and *accomplice*, we provide *@MetaphorMagnet* with positive and negative exemplars for each in the context of over 1000 actions in all. Thus, the ideal accomplice is *loyal* and *tight-lipped*, but the worst is *treacherous*, perhaps even an *undercover cop*. These ideals (and anti-ideals) are used directly in various ways by the system and its kindred bots on Twitter, as in this piece of faux advice from the *@BestOfBotWorlds* bot:

> *from "The 9 Habits of Highly Organized Crooks":*
> 1. *If you're gonna embrace lawlessness, embrace lawlessness with lovable rogues.*

While negative exemplars suggest parodies of religious moral proscriptions in the vein of the 10 commandments:

> *Commandment XXVI: Landscape gardeners, thou shalt NOT create lawns with unholy laborers that doth toil on the Sabbath.* #ThingsMosesNeverSaid

These are the low-hanging fruits of a moral value system, but the larger purpose is to use such idealized role fillers in stories that exemplify one's good (and bad) choice of roles. At their simplest, such stories bring together a protagonist and antagonist to vividly play different roles in the same event. To lend its moral mini-stories an Aesopian quality, *@MetaphorMagnet* first generates a pair of property-based animal metaphors for its chosen role-fillers. As animals operate primarily on instinct they represent the Freudian id, which is framed (and tamed) by the moral tone of the story. A story such as the following is then tweeted in two parts:

> *A charming horse once demanded a bribe.*
> *A criminal crab then thought "A murderer like me needs a victim like this for my assaults."*
>
> *"The best victims are unpopular politicians," thought the criminal crab. So the crab decided to commit an assault against the charming horse.*

@MetaphorMagnet draws upon a database of stereotypical properties to know that politicians are typically charming, and uses the Google n-grams to suggest "*charming horse*" (freq=411) as an apt animal metaphor for *politician*. It uses the alliteration of *criminal* and *crab*, and knowledge of the negative lexical affect of both words, to suggest "*criminal crab*" as an apt animal metaphor for *murderer*. It then finds a unifying event in which each can jointly participant, in this case in the roles of *victim* and *murderer* respectively. These are simple one-act stories with a clear (if subjective) moral, concisely packed into the most resonant of forms. So we now consider longer-form narratives of sequential character interactions bookended by a moralistic message.

For Every Action, A Moral Reaction

Actions with a moral dimension often elicit a moral action in response. The challenge of creating an engaging moral

story is to find a novel but plausibly familiar drama in this moral *tit-for-tat*. For the panoply of such stories defines the space of the moral imagination, so the more new stories we can tell, the more of this space we may learn to navigate. We begin then with our core assumption that actions with a moral dimension are likely to elicit moral reactions in turn.

We define a moral reaction triple as a sequence of interlaced actions between protagonist X and antagonist Y:

1. X performs an initial action on Y
2. Y reacts to 1 by performing an apt action on X
3. X reacts to 2 by performing a new action on Y

For instance, if (1) *X exploits Y* then (2) *Y distrusts X,* so (3) *X alienates Y*. This moral triptych is a tiny drama in its own right, comprising simple actions and reactions that reflect our time-tested insights into human behavior. More complex dramas can be now constructed by tiling adjacent triples together, so that their actions or reactions overlap. Two triples T_i and T_j are adjacent if the last action of T_i is the first action of T_j, yielding a T_i:T_j tiling of five actions. In effect, these triples are not so much *master plots*, in the sense of Cook's (1928) PLOTTO, but *plot segments* that are intended to be joined-up as one lays the tracks of a train set. Though an AI system might conceivably acquire these plot segments automatically from a large, annotated story corpus, we take it as our task here – as in Cook's PLOTTO system – to lay out this core set of combinatorial elements manually, to ensure the moral coherence of their linkages. Thus, so as to knit together more coherent plots, we favour action-reaction-action triples over action-reaction pairs.

We name this approach ***Scéalextric***, from the Irish *Scéal* (meaning *story* and pronounced *scale*) and the brand-name *Scalextric*, a racing car simulation in which hobbyists build complex race tracks from composable track pieces, so as to then race miniature electric cars upon these racetracks. The more varied the track segments that hobbyists have at their disposal, the more dramatic the racing narratives that can emerge from their simulations in miniature. For instance, only a track segment that allows two racing lanes to cross-over will ever allow for the narrative possibility of two cars crashing into each other. Hobbyists seek out a large variety of track segments with the widest range of affordances, to build complex tracks that can give rise to satisfying race simulations (that is, race *narratives*). By analogy, the plot triples of *Scéalextric* are the track segments from which a satisfying moral narrative can be constructed, but these too must afford diverse moral interactions between characters. *Scéalextric* at present comprises over 2000 plot triples (on the same order then of PLOTTO's 1500 master plots). Given a random start action, one can chart a path through the space of moral possibility by first choosing from all the triples that begin with this action, and by then choosing amongst all the possibilities to link this triple to others, and so on until a satisfactory ending action has been selected. But what constitutes morally satisfying starting and ending actions for our automated tales? *Scéalextric* define a moral bookend for each action in its repertoire: if a story begins with the action A, then an opening bookend defined for A is used to start the story; if a story ends with an action Ω then one of the closing bookends for Ω will close out the story. The bookends for A and Ω provide the moral frame in which to view the actions that link A to Ω. Consider this short story, generated by *@MetaphorMagnet* in 7 tweets:

The Neoconservative And The Conservative:
Aesop's lost tale of a sinister cat and a political gorilla

The gorilla stood out for the cat in a field of weak candidates.

So at first, the sinister cat voted for the political gorilla in the election. But the gorilla's flaws became all too apparent to the cat.

So the cat campaigned vigorously against the gorilla. So the gorilla intimidated the cat with threats of violence.

Yet the cat delivered a crushing defeat to the gorilla. But the gorilla bribed the cat to play along.

So in the end, the cat forgave the gorilla for all its sins. That's the way it should be: only forgiveness can wipe the slate clean. ***The End***

The digested read*: The sinister cat embraced conservatism, while the political gorilla believed in conservatism.*

The above story demonstrates a number of qualities – and a crucial issue that has yet to be resolved – with *Scéalextric*. First, it is rendered as a chain of idiomatic text fragments rather than as a chain of conceptual primitives. An action lexicon maps from primitive actions (e.g. *X disappoints Y*) onto a range of idiomatic forms, such as the rendering "*X's flaws became all too apparent to Y*" that is chosen above. Second, the conceptual characters in the story – the *neoconservative* and the *conservative* – are sourced from the knowledge-base of stereotypes that underpins the workings of *@MetaphorMagnet* (see Veale, 2014, 2015). So it is the stereotypical actions associated with these types – e.g. that conservatives vote for other conservatives, and that neoconservatives seek to win votes – that allows *Scéalextric* to choose an apt starting triple for the story (namely: *Y impresses X*; so *X votes for Y*; but *Y disappoints X*). The discourse linkages "so", "but" and "yet" are obtained from an action graph that models the relationship of successive actions to each other, where *but* and *yet* each indicate surprise and *so* indicates unsurprising consequence. Use of the latter results in a linear narrative without surprises, and use of the former introduces twists and turns into a plot.

In our analogy of plots as racetracks for action, the plot triples that introduce *but* and *yet* links are the curved pieces of track (perhaps hairpin bends if used in rapid succession), while *so*-links allow *Scéalextric* to build a straight piece of unbending narrative. We have yet to empirically determine

the optimal balance of *so*'s and *but*'s for a *Scéalextric* plot, but leave this task for now as the subject of future research. As we plan to open up the *Scéalextric* system and release its databases of actions, triples, idioms and discourse links publically, we expect others may also have insights on this.

But the above also demonstrates a weakness in the track-laying approach to narrative, for the moral of the story can only be judged relative to the desirability of the final act, as it is this that cues up a moral framing for the story. But this heuristic of *all's well that ends well* is clearly shown to be deficient in this case, as forgiveness – a desirable moral act – is here achieved via immoral bribery. What is needed is a global perspective on the morality of the tale as a whole. To this end we intend to integrate notions from classical narratology, such as Propp's structuralist view of narrative in terms of character functions (such as *hero* & *villain*, but also *false hero* & *helper*) and their effects on a plot (see Propp, 1968). Propp's morphology of the folk tale has been shown to offer an effective framework for the automated generation of tales by a computer (e.g. see Gervás, 2013). By making character function an integral part of the triple-selection and plot-laying process, we believe we can give *Scéalextric* an ability to properly track the progress of hero and villain through a tale, to thus ensure that the resulting plot conforms to our expectations of their moral purposes.

Concluding Thoughts: And the moral is …

Stories serve a pivotal role in the development and active functional use of our moral imaginations. As Freud once put it, "*The virtuous man contents himself with dreaming that which the wicked man does in actual life*." Functional morality – a form of moral thinking that can influence our actions and our thoughts about situations that have yet to be encountered, must be contrasted here with operational morality, the ethical restrictions placed on the actions that a system can take, and indeed on the thoughts it can explore, by a concerned engineer (see Wallach & Allen, 2008). The latter yields rigid systems that become brittle in the face of unanticipated scenarios; the former, though more desirable, requires us to give our computers a moral imagination as a sandbox for imagining the possible consequences of their actions (and indeed their inactions) in novel situations.

We have argued here that the ability to generate and understand stories with a moral is key to the development of this moral imagination, both in humans and in machines. Starting from a value-free model of the world that employs lexical sentiment as a crude substitute for moral judgment, we have shown that machines can employ the moralizing style of human orators to good effect, to pose interesting moral dilemmas for humans that it itself cannot appreciate.

As its next step along the path to functional morality, we gave *@MetaphorMagnet* a rather simple value system, by characterizing the roles that may participate in the various actions in its knowledge-base, and by offering positive and negative exemplars – ideals and anti-ideals – against which it might evaluate the morality of any role in any action. But this value system considers individual actions in isolation, rather than the *tit-for-tat* sequences that often emerge when inter-personal actions are freighted with moral dimensions.

So we have broadened *@MetaphorMagnet*'s horizon to include triptychs of back-and-forth actions between pairs of characters. This approach to narrative construction from reusable plot segments – effectively action-level n-grams – holds out the promise of delivering long-form stories with both a twist and a moral in the tail. As we open *Scéalextric* to others and make its database of actions, triples and plot linkages publically available, we welcome a new chapter in the creation of AI systems with a moral imagination.

Acknowledgements

This research was supported by the EC project *WHIM: The What-If Machine*. *http://www.whim-project.eu/*

References

Baker, C. F., Fillmore, C. J. and Lowe, J. B. 1998. The Berkeley FrameNet project. In *the Proceedings of COLING-ACL'98,* Montreal, Canada pp. 86-90.

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

Cook, W. W. 1928/2011. *PLOTTO: The Master Book of All Plots*. Tin House Press re-edition of the 1928 first edition.

Fauconnier, G. and Turner, M. 1998. Conceptual Integration Networks. *Cognitive Science,* 22(2):133–187.

Freud, S. 1933. *New Introductory Lectures on Psychoanalysis*. London, UK: Penguin Freud Library 2.

Gervás, P. 2013. Propp's Morphology of the Folk Tale as a Grammar for Generation. In *Proceedings of the 2013 Workshop on Computational Models of Narrative*, Dagstuhl, Germany.

Hunter, D. 2008. Teaching and Using Analogy in Law. J*ournal of the Association of Legal Writing Directors*, Vol. 2.

Koestler, A. 1964. *The Act of Creation*. Hutchinsons, London.

Propp, V. 1968. *Morphology Of The Folk Tale* (second edition). University of Texas Press.

Veale, T. and D. O'Donoghue. 2000. Computation and Blending. *Cognitive Linguistics*, 11(3-4):253-281.

Veale, T. 2014. Coming Good and Breaking Bad: Generating Transformative Character Arcs For Use in Compelling Stories. *Proceedings of ICCC-2014, the 5th International Conference on Computational Creativity, Ljubljana, June 2014*.

Veale, T. and Valitutti, A. 2014. A World With or Without You* (*Terms and Conditions May Apply). *Proc. of the AAAI-2014 Fall Symposium Series on Modeling Changing Perspectives: Reconceptualizing Sensorimotor Experiences. Arlington, VA*.

Veale, T. 2015. Game of Tropes: Exploring the Placebo Effect in Computational Creativity. *In the Proceedings of ICCC-2015, the Sixth International Conference on Computational Creativity,* Park City, Utah, May 31-June 3, 2015.

Wallach, W. and Allen, C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. London, UK: Oxford University Press.