Metaphor, Blending and Irony in Action: Creative Performance as Interpretation and Emotionally-Grounded Choice

Tony Veale
School of Computer Science
University College Dublin
Dublin, Ireland.
Tony.Veale@UCD.ie

Philipp Wicke
School of Computer Science
University College Dublin
Dublin, Ireland.
Philipp.Wicke@UCDConnect.ie

Abstract

Metaphor is a powerful tool in the performer's tool box, not least because it can operate at several levels at once. As our linguistic metaphors deliver rhetorical flourishes, conceptual metaphors change the way we see the world, while our metaphorical *body language* allows us to take postures and stances that are both figurative and literal. Each kind of metaphor is another expressive choice that creative performers can make to convey their meanings. This choice sits at the heart of what it means to be a creative agent, for agents who lack choice and do exactly as they are told cannot exercise genuine intent for their actions. We shall explore how interpretation is wrapped up with choice to appreciate and to achieve emotional creativity in a system for generating and enacting automated tales. We explain here how three levels of description – linguistic, conceptual and physical – are integrated in a framework that motivates the use of metaphor or irony as a creative choice, by robot performers aiming to go beyond a predetermined script.

Interpretation Makes It So

Hamlet reminds us that "there is nothing either good or bad but thinking makes it so." We form our judgments, creative or otherwise, by viewing situations and events through the lens of interpretation. It is a capacity for interpretation that allows actors to deliver a creative performance even when their actions and their lines are dictated for them. Indeed, it is interpretation that allows scripted performers of any kind to do more than obediently follow the script they are given, and to bring something of themselves to their work. It is in the gap between interpretation and execution that creativity can take root and blossom, even for the most scripted roles.

In the context of a programming language such as *Java* or *Python*, interpretation is wholly deterministic, and offers no avenues for free choice. In contrast, when one interprets natural language, there is always some scope to equivocate, to exaggerate, or to place an emotional spin on events. This is the view of interpretation that we set out to explore here: interpretation as a construal that understands why one word or action leads to another and produces the feelings it does. When performers interpret a script to create a performance, they go from a conceptual appreciation of cause and effect to an emotional appreciation of what should be expressed. Unlike the literal interpretation of a programming language or a rigid script, which brooks no uncertainty, no ambiguity or no variation for its own sake, creative interpretation sees scripts as elastic starting points, not frozen end points. With the situational awareness to appreciate when departure from the script is warranted, performers can choose to interpret a directive with under- or over-statement, metaphor or irony.

The dichotomy between freedom of choice and no choice at all is an extreme one, but there are more subtle dilemmas. There is, for one, a real distinction between true choice and empty choice, or choice for its own sake. At any given time an agent may have several valid responses available to it, thus offering it an opening for a creative choice. If an agent chooses randomly from these possibilities, or if it chooses so as to avoid repetition and foster variety, it is making a hollow choice based on form without meaning. This is the essence of what is often called *mere generation* (Veale 2012; Ventura 2016): the glib production of well-formed outputs that are valid only because they obey the rules of the game, not because they have an inherent value that is appreciated by the generator itself. Mere generation uses generic rules to make specific choices, but leaves the interpretation of the specific ways in which the rules are instantiated to the user.

The rules of a merely generative game are themselves a script into which formal choices like these can be baked in from the outset. A disjunctive script that says *do this or this or that* can be just as rigid as one with no disjunction at all. Choice without interpretation leads to empty variations that thwart self-evaluation and lack creative force. We critique such an approach to story-based performance here, one that uses disjunction to achieve diversity of output without truly appreciating the meaning of that diversity. With this as our baseline, we model what it means for a performer's choices to be driven by an emotional understanding of the script. A performer should choose to react to a narrative event with a metaphorical exaggeration, or even an ironic response, not just because this is a possibility, but because it enhances the telling of the story to react this way. This will require us to insert an emotional layer between the conceptual layer of the system – its plots, actions and model of cause and effect – and its expressive layer of physical gestures and spoken words. This sandwich of distinct layers will allow performers to make informed choices that are grounded in context.

Interpretation is the missing ingredient that fills the gap between mere generation and intentional creation. It is key

to a producer's efforts to make informed choices, just as it is key to a consumer's efforts to ascribe value to any outputs. Interpretation recognizes the value of a departure from the scripted norm, but interpretation is more than recognition. To appreciate the difference between these related notions, consider the following example of an email communication from the first author to the second, as sent via *Gmail*:

> We went to see "Tenet" last night. Our brains are still bent out of shape. But we plan to see it again last week, so we'll understand it eventually.

The joke, such as it is, relies on the specifics of "*Tenet*", a time-travel movie in which mysterious agents use reverse entropy to go back in time. The statistical language model used by *Gmail* (Chen, Lee, and Bansal 2019) to predict future texts and flag possible errors in past texts spies a rather improbable word choice in "again *last* week," and it dutifully underlines "last" in blue to signal its recognition of this departure from the norm. But the model does not "get" the joke – it sees no link between "Tenet" and the future-looking use of "last week"– and so cannot interpret it as irony. This is a consumer view of a producer's attempt at wit, but what of the producer's perspective? Let's choose a more normative situation that relies only on general world knowledge:

> I have never married, but I have had a few near misses.

To re-package this metaphor as a joke, we might replace the high-probability word choice of "misses" with the phonetically identical, but much more improbable, "Mrs":

> I have never married, but I have had a few near Mrs.

The replacement yields a recognizable departure from the norm that a language model like *Gmail*'s can easily detect, but mere recognition of this departure is not enough for wit. We could, for instance, have replaced "misses" with "kisses" or "fishes" to achieve the same low-probability punning surprise, but such a replacement would make little or no sense. It takes interpretation to appreciate the logic of "Mrs" in this context, since only "Mrs" produces the same humorous kick. Ideally, the interpretation of the producer will mirror that of the consumer, to predict the surprise *and* its final resolution.

The rest of the paper explores how an interpretative layer can be inserted into an existing computational framework for the embodied performance of machine-generated stories. We begin by considering this existing system as a baseline, to ascertain the degree to which its choices are disjunctive but empty, or interpretation-driven and potentially creative. We aim to shift its workings closer to the latter by basing its decisions on an emotional understanding of a story's events. Disjunctive choice can be a useful source of plot variation if the disjunction is motivated by the emotions of the story's characters or an audience's attitudes to them. We show how a balance is realized by allowing audiences to influence the plot as they reveal their own feelings for a story's characters. But what matters is a performer's choices and how they are interpreted by the audience. As embodied actors that bring a tale to life with their gestures and spatial movements, these choices must be seen as coherent for the plot. We present a crowd-sourced evaluation that shows this is indeed the case.

A Critique of Pure Disjunction

The *Scéalextric* framework of (Veale 2017) adopts a *story grammar* approach to plot generation. Its tales are assembled in a click-&-build fashion from a large set of prefabricated plot segments of three successive actions apiece. Inspired by the *Plotto* framework of (Cook 1928) – one of the oldest structuralist approaches to systematic story construction – a *Scéalextric* plot is first assembled by linking a series of plot triples end-to-end, or by recursively refining a single over-arching triple in a top-down fashion to flesh out a narrative arc. Cook listed almost 1,600 plot triples in his 1928 book, while *Scéalextric*'s stock numbers over 3,000, each crafted from a set of 800 verbs that relate the generic characters A and B. As this genesis in Cook's approach will testify, *Scéalextric* does not employ a particularly novel approach to plotting. Rather, its appeal lies in its scale, modularity and openness. A skeletal plot is easily rendered in English with a large set of idiomatic renderings that map actions from the logical to the text level, via forms that include idioms and metaphors. The system's data-rich modules are open for all to use, and new modules – to e.g. add dialogue, as in (Wicke and Veale 2020), or to append a moral – are easily defined.

In top-down mode, the initial triple is used to give shape to a story. In end-to-end mode, a new triple is added to the growing plot if it shares a connecting action, and if no loop or repetition results from its addition. So, the choices made during plot assembly are formal ones, decided on the basis of compatibility with the story grammar rather than for any semantic or narratological reasons. These wholly structural decisions require no interpretation of the evolving plot, and so can be considered merely generative disjunctive choices. As such, *Scéalextric* exhibits broad generativity but little or no true creativity in the assembly of its plots or in the final rendering of those structures as idiomatic English stories.

So where does the creativity, if any, arise in this system? As described in (Veale 2017), *Scéalextric* goes beyond the purely disjunctive to choose the characters that will fill the A & B positions in its plot skeletons. A database of familiar characters from fiction and history, called the *NOC List*, is used to fill each role with a recognizable personality, so that each character's extensive backstory (as stored in the NOC) can be woven into the rendering of the tale. For instance, if the plot calls for A to insult B, the idiomatic rendering of *insult* offers a generic account of the offense, but the NOC allows specific negative details of the target to be aired too. The NOC also allows metaphors and similes to be coined on the fly, so that e.g. *Richard Nixon* insults *Bill Clinton* by likening him to *Pepé Le Pew*, or insults *Frank Underwood* by comparing him to *Keyser Söze*, thus winkingly breaking the fourth wall at the same time (as the NOC knows that each character was portrayed on screen by the same actor).

Crucially, NOC characters are chosen for their suitability to specific actions in the plot. If the plot turns on a *betrayal*, a sneaky character is chosen; and if one character must heal another, a doctor is chosen. To achieve a measure of wit, A & B are instantiated as a pair, and this is where *Scéalextric* makes semantic choices that are guided by interpretation. As outlined in (Veale 2017), character pairings are chosen to exhibit a mix of appropriateness and incongruity, or what

humour theorist Elliott Oring calls *appropriate incongruity* (Oring 2011). As in the *last/next* and *misses/Mrs* oppositions explored in the introduction, a good pairing produces a reassuring surprise, an apparent mistake that only makes sense on closer examination. Two NOC entities are paired if they are linked in the popular imagination – perhaps they were portrayed by the same actor, or created by the same author, or belong to the same group, or share some key properties – and there is also some common-sense bar to their union, such as that one is historical and the other fictional, or they belong to different fictional franchises or historical periods. *Scéalextric* might pit Alan Turing against Sherlock Holmes, or make business partners of Ada Lovelace and Steve Jobs. The effects of each high-friction pairing percolate through the rendering of the tale as a whole, reminding audiences of the appropriate incongruity at the story's core, as aspects of each character – properties, clothing, or physical settings – are integrated into the surface renderings of plot actions.

Scéalextric has been used as a generative basis for other story-telling systems. Wicke and Veale (2018) used robots to enact its tales with physical gestures and spoken dialogue. Veale, Wicke and Mildner (2019) subsequently built a model of performance, called *Scéalability*, around this generative core, allowing for tales to be physically enacted by a cast of robots and smart devices. Anthropomorphic robots move about a stage as they act out the central roles of a story, while an omniscient narration is voiced by an *Amazon Echo/Alexa*. Additional dialogue is layered over the textual rendering of each story, so that the robots have apt lines to speak as they move and gesticulate. The dialogue module that generates this spoken script is powered by simple disjunctive choice: for each of the 800 actions in the *Scéalextric* plot vocabulary a set of dialogue fragments for A and for B is defined, and the actors choose randomly from this set for a given action. This simple approach is especially effective in stories with more than two characters. Since roles such as A-spouse and B-friend cannot be physically enacted with just two robots, we know of their actions only from the narrator, and from the commentary of the main actors as they react to events.

The physical actions of the *Scéalability* robots are also determined by simple disjunctive choice. For each of the 800 possible actions in a *Scéalextric* plot, a set of motor scripts is associated with the A and B roles, and the robot performers are free to choose which script to execute. As in the dialogue, these choices are always constrained by the needs of the current action. No consideration of past actions, and of how they influence an audience's interpretation of the current action, is brought to bear, and no freedom is given to depart from the script. This prevents the performers from interpreting their scripts, to decide that a certain plot point needs to be emotionally heightened with metaphor or irony. We aim here to remedy this lack of interpretative freedom.

Plot Disjunction at Time of Performance

A story grammar is essentially a causal graph of actions and their consequences. A "walk" through this graph, whether a random walk or a goal-directed journey, yields a single path and a single plot line. Branching points in the causal graph present choices that are resolved at the time of the walk, not at the time of the story's performance for an audience. But this need not be the case: if branch-points are inserted into the plot, turning it from a line into a tree, those choices can be resolved latter, perhaps with the help of the audience.

When a story generator makes these choices for itself, by treating each branch point as a purely disjunctive choice, it simply explores the space of possible stories without regard for the emotions of its characters and those of an audience. Interpretation is supposed to offer insights on such matters to the performers, but by this time the tale has been written. By allowing choices to be made at the time of performance, an interpretation of what has gone before in the narrative can shape the course that the performers will take. Indeed, the performers can involve the audience in their decisions, so that they make choices that seem emotionally plausible.

We can use the word "script" to denote the sequence of actions to be followed by a performer, or a body of code to be executed on a machine. As we have seen, each involves a different idea of "interpretation." While the latter brooks no flexibility, no metaphor, and no loose readings of the text, it does allow for conditional *if-then-else* branching structures. To support performance-time decisions regarding plots, we incorporate both senses of "script" into *Scéalextric* stories. It is a simple matter for a story-grammar to generate *if*, *then* and *else* markers in its plot lines, and to recursively expand different plot continuations after a conditional branch point. The resulting plot is still a linear sequence of symbols, but, like a computer program, it is executed by its performers in a dynamic, non-linear fashion. When robot actors resolve a branch point for themselves, they can use the interpretative, emotion-based mechanism we present in the next section. Or they can ask an audience to provide an interpretation for them, falling back on their own logic when none is offered.

Certain plot actions represent dramatic choice points in a story, as when, for instance, A considers forgiving B for an earlier offense. It is at these points that the story grammar obtains maximal benefit from a disjunctive turn in the plot, as these emotion-laden turning points should also elicit an emotional response from the audience. To elicit a response, the performers explicitly ask the audience for their input at these junctures, by e.g., asking "Should I forgive this guy or not?" To register this response, if one is forthcoming, a video camera is used to capture the facial expressions and the hand gestures of audience members. Since the robots themselves use gestures to convey their emotions, it seems fitting that the audience likewise joins in the performance.

The robot whose character is to perform (or not) the given action pauses, turns to the audience, and poses its question. The system's camera is constantly trained on the audience, but its video feed is only examined in the moments after the question is posed. The robot also gestures to signify that it awaits an answer, but is capable of carrying on without one. Two distinct neural networks examine the same images: the first (Cao et al. 2018) scans the video for hand gestures, returning a label, a bounding box and a confidence score; the second (Goodfellow et al. 2013) scans the image for facial emotions, likewise returning a label, a location and a score (see Fig. 1). The performer makes its decision on the basis of both data sources as weighted by their confidence scores.

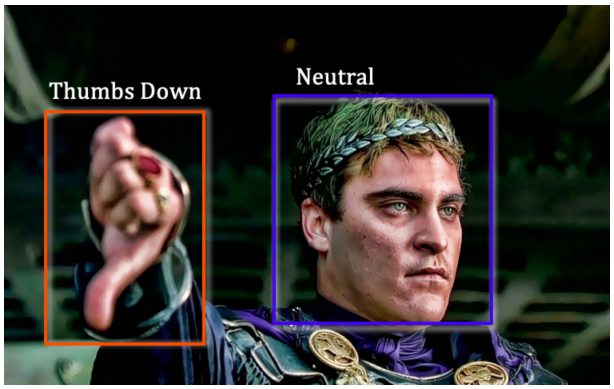

Figure 1: Gestural (red) and facial (blue) analysis as applied to a sample image. The video feed is scanned for hand and face signals in the moments after a robot poses a question.

At present, the system considers thumbs-up as a clear "yes" and thumbs-down as a clear "no." If the feed presents data from several audience members, the most confident results are used. So hand gestures provide a binary yes or no, while facial expressions supplement this answer with an emotion. This emotion (a sad face, an angry face, a happy face, etc.) is used to decide the question if no hand gesture is detected. If both are present, a contingency table is used to resolve the matter, especially in cases of mixed signalling, as when e.g. a user gives a sad thumbs-up or a happy thumbs-down.

Once More, With Feeling!

Tapping into an audience's reactions in this way allows the actors to *borrow* their emotional interpretation and make it their own. However, when those reactions are not apparent, the performers must arrive at their own interpretation of the current state and the actions that led to it. In fact, performers must do this for themselves anyway, for all of the other states that do not correspond to an explicit branch point in the plot.

The *Scéalextric* story-grammar does not permit repetition of the same action within the same plot, and so each action is unique, and denotes a unique juncture within each story. However, the current state of the narrative is more than the current action, and must include all the expectations that we carry into it from past events. Those expectations determine the extent to which the current action is surprising, and the extent to which characters feel shocked or disappointed by unfolding events. We might, for instance, expect characters that are shocked by the current action to react with greater emotional force than ones who see it as a natural outcome.

We must first characterize the emotions arising from a single action, for both its agent and the patient it affects. We can then quantify the halo of emotions that carry over from earlier states and add to the current emotional load. The range of emotions that we can distinguish – from a set of basic or pure emotions to complex blends of these primary colours – is large (Ortony, Clore, and Collins 1988), but a survey of the 800 plot actions suggests eight that are most useful: *respect, disrespect, inspiration, disappointment, support, aggression, attraction* and *repulsion*. Specifically, we want to quantify the degree to which one participant to an action may feel respected or disrespected, inspired or disappointed, supported or attacked, attracted or repelled, and calmed or aroused. We will quantify by degrees, on five parallel scales:

disrespected+++	.....▽.....	respected+++
disappointed+++	.....▽....	inspired+++
attacked+++	...▽.....	supported+++
repelled+++	.....▽.....	attracted+++
calmed+++	.....▽....	aroused+++

For every action in the *Scéalextric* vocabulary, and for each A and B role of that action, we mark the expected position of the character filling the role on each of these five scales. For example, for the action A *worshipped by* B, we record that A feels *respected*+++ (the highest degree of respect) and B feels *inspired*+++ (the highest level of inspiration). Conversely, for the action A *betrayed by* B, we record that A feels *supported*- - - (the lowest level of support) while B feels *inspired*- - (a low level of inspiration, but not the lowest). We do this on five scales for each role in all 800 actions. The arousal scale marks the intensity of each response. For instance, one is more aroused when one hates than when one dislikes, or when one worships rather than merely admires.

The first four dimensions are emotionally charged, since they mark out emotions with a positive or negative valence. The mean value of these dimensions thus provides the overall valence of an action from the perspective of a given actor. The fifth is not charged in this way – one can be as aroused by hate as by love – but it does signify the energy with which a feeling is experienced. Valence quantifies the impact of an action on an actor, while arousal suggests the scale of the actor's dramatic response (Kensinger and Schacter 2006). A high-arousal feeling calls for a dramatic, high-energy gesture; a low-arousal feeling calls for a more subtle enactment.

To capture the emotional inertia of a character, we need more than the mean valence of their feelings at a given time. A weighted average of the aggregate valence of a character's feelings from one action to the next – giving 50% weight to the current action – allows us to smoothly track changes in a character's perspective over time. When this inertial valence undergoes a significant shift, to the positive (a sudden boon) or the negative (a sudden disappointment), this indicates a macro-level change that merits a macro-level interpretation.

In a sense, the horizontal scales mirror the physical stage, since each response can move actors closer together or further apart. A and B maintain an emotional distance to each other that grows or shrinks with each new plot event. While there is no quantitative difference between *supported*- - - and *attacked*+++ there is a qualitative one: the former reflects a failed expectation of support, while the latter incorporates no prior expectation of aggression. A character who is betrayed is not just figuratively attacked by another; they have their expectations of support dashed at the same time. Likewise, *attacked*- - - reflects a lack of expected aggression and a high level of support, as when A *surrenders to* B. So this notation permits us to take a wholly quantitative view of the emotional effects of an action, while also allowing us to encapsulate a qualitative sense of our surprise at those effects. This will prove especially useful when we consider irony.

We can consider each parallel scale in isolation or in the aggregate. When considered in isolation, we can compare the current scalar settings to those of the previous action, to quantify the emotional shift wrought by the current action. When this current action comes as a surprise, a plot twist of sorts, we can expect significant jumps on some or all scales. Plot twists in *Scéalextric* are woven into its story grammar, which also dictates the use of 'but' or 'then' to link actions. As such, its twists are purely formal products of disjunctive choice that are not interpreted emotionally by the generator. Yet, as the grammar tells the generator to insert *but* or *then*, an emotional layer can explain why this must be so, in terms of how the characters experience this turn of events. This is what it means for the system and its performers to interpret a story action in the light of past events. It is on the basis of this interpretation – the emotional shift from one event to another – that performers can choose to react figuratively, to obey the scripted norm or use metaphor or irony instead.

When should a performer choose a figurative response? Without interpretation, metaphor is just another disjunctive choice. With interpretation, a performer can reason that the scripted response is inadequate for the context in which the action is being performed. The standard script is inadequate when the feelings evoked in the moment are more intense than the current action, viewed in isolation, would suggest. Metaphor, or indeed irony, is an opportunity to recalibrate and fine tune the performer's responses to suit the moment. Suppose A has shown favour to B in some way, perhaps by promoting B, or sharing a story with B, or confiding in B, and B responds by insulting A. Viewed in isolation, this act of repudiation should make A feel quite *disrespected* (++), and even somewhat *attacked* (+). The dialogue model will suggest a scripted response to A that mirrors these feelings. However, when seen in the context of its previous actions, A should feel even more disrespected (+++) and feel all the more attacked (++ or even +++). The most apt response for A, then, is not to act as if insulted, and to speak and gesture accordingly, but to react as though physically attacked. That is, A should speak the lines of an attack victim, and act out the gestures of one who is under attack (e.g., "Get off me!" spoken while extending the arms and stepping backwards).

Taken together, the words and gestures of A and B are no longer a performance of the single action A *insulted by* B. Rather, the result is a conceptual blend in words and actions (Fauconnier and Turner 2002): B acts out an insult while A acts out an attack. This captures the truth of the situation as seen by A, but it also produces a novel blend that adds diversity to the performance. Actions interpreted and enacted in context lead to more varied enactments than they otherwise would. An ironic response emerges in much the same way, though the ensuing blend is more granular. Consider an action, A *stands up to* B, that has A feeling uninspired (*inspired–*). While *inspired–* is comparable to *disappointed++* in scalar terms, since each causes the same shift on the same scale, it also encodes a failure to be as inspired as one expects. Irony is a playful response to this failure of expectations, insofar as it signals the failure while pretending that it has not occurred, thus highlighting the gulf between expectation and reality (Garmendia 2018). To pretend in a performance context, a performer can invert the valence of the expectation, e.g., to obtain *inspired++* from *inspired- -*, and then act out an action associated with the inverted emotion.

For instance, A can ironically perform A *bow down to* B instead of A *stand up to* B, by actually bowing to B, while nonetheless speaking the lines scripted for A *stand up to* B. The clear friction between this dialogue (for example, "I've had enough of you!") and the oddly respectful gesture tips the wink to the audience that all is not as it seems. We use the gestural channel to carry the irony because it is a much more suggestive, and far less direct, carrier of meaning.

These approaches to irony and metaphor are compatible, and play well together to create an ironic overstatement. For instance, if metaphor is used to map A *bow down to* B onto A *worship* B, to exaggerate the extent to which A looks up to B, the irony of acting out the supplicant gestures of A *worship* B while vocalizing A *stand up to* B is sharper still. This remains a relatively safe means of incorporating irony into a performance, as the spoken dialogue keeps the action moored to the literal basis of the story. Gestures are often more subtle than words, and offer more plasticity to an actor.

Robots offer quirkier modes of expression, such as flashing LEDs, that are less intuitive than either words or gestures (Häring, Bee, and André 2011). Indeed, gestures are a vital part of embodied communication (McNeill 1992), and many have the same deep, conceptual roots as words. Still, a fixed set of task-specific gestures is often defined for robotic performances (Wilcock and Jokinen 2013; Csapo et al. 2012), but even these bespoke gestures are subject to cultural variability. Speakers of the Aymara language, for instance, refer to future events not by pointing ahead but by pointing *behind* (Núñez and Sweetser 2006). It pays, therefore, as far as it is practicable, to rely less on iconic gestures that are culturally rooted (such as kneeling to propose) and more on schematic movements that have a greater claim to universality.

A performer's gestures should always rhyme with their words, unless irony is used to create a knowing dissonance. The approach to metaphor and irony presented here works primarily at the conceptual level of plot actions, and so the choice of words and gestures follows from this. Metaphors arise when a plot action is mapped to another, semantically similar action that more intensely evokes a certain emotion. This intensification explains why metaphor is an asymmetric form of comparison that marries similarity to directionality. Conversely, irony is achieved when one action is mapped to a semantic opposite that evokes the inverse of an emotion. So, irony relies on opposition rather than similarity, and on inversion of emotions rather than their intensification. Still, a mix of irony and metaphor can use opposition, similarity, inversion *and* intensification to produce a satirical effect.

In each case, however, after a mapping between actions is achieved at the plot level, the corresponding dialogue and gestures are chosen because they happen to be associated *a priori* with the given actions. While these purely disjunctive choices are driven by emotional choices at the action level, they can also be grounded in an emotional interpretation. Consider how a gesture is chosen to enact a specific action. Many gestures are pantomimic and culture-specific, but other movements, such as relative motion between per-

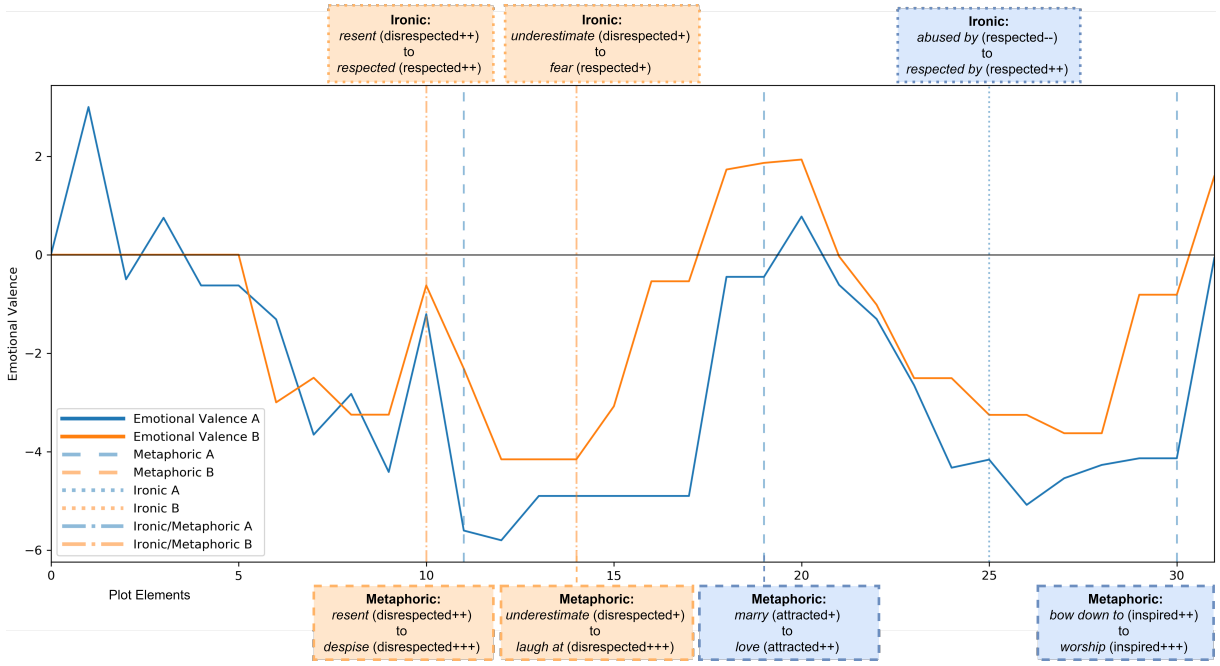

Figure 2: Interpretative decision points in a story with 31 plot actions. Annotations at top explain ironic interpretations. Boxes at bottom explain metaphorical exaggerations. Blue lines track the inertial valence of character A; orange lines track that of B.

formers, are more subtle and less dramatic, but just as communicative (Wicke and Veale 2021). The arbitrariness of many gestures in cultural and dramatic terms makes it appealing to simply define gestures as black-box scripts for different actions. But we can also annotate gestures on an emotional level, using the same emotion scales that are employed for plot actions.

For example, the bowing gesture is now annotated with *respected*++ for the actor in front of which it is performed, and annotated with *inspired*+ for the actor that performs it. It is vital that we separate each gesture from the action that it may embody, so as to annotate each one on its own terms. This ensures that the emotions we associate with a gesture actually reflect the audience's reactions to this gesture, and not our desired reaction to the action it means to enact. As this is a tricky knot to unpick, we plan to crowd-source the emotional annotations for each gesture in an empty context. For the present, we annotate each gesture ourselves, so that the performers can use these annotations to select the most appropriate gesture to perform for any given action. In this way, robot performers can choose the gestures that reflect the emotions of a scene as *they* interpret them in context.

Opportunities for Metaphor and Irony

Valence and arousal are key elements of suspense (Delatorre et al. 2016). They enable us to keep listeners on the edge of their seats before arousing them with a sudden plot twist. At key points, the inertial valence of a character's perspective can flip from positive to negative, or vice versa. When a substantial shift in inertial valence occurs at the current action, it will have been brewing for some time. When this shift exceeds a fixed threshold Δ, an actor might mark the shift with an exaggerated response that goes beyond the script. Fig. 2, which tracks the inertial valence of characters that fill the A (blue line) and B (orange line) roles in a *Scéalextric* plot of 31 actions, highlights shifts that support metaphor and irony.

These shifts are marked by vertical lines in Fig. 2: dashed lines mark those supporting a metaphorical response, dotted lines mark those that support an ironic response. The plot in Fig. 2 offers three shifts where irony is supported (upper annotations) and four where metaphor is a strong possibility (lower annotations). Irony is a valid response when the fall in a character's inertial valence exceeds the threshold Δ, signifying a failure of expectations. Irony is also supported when an action is explicitly tagged with a failed expectation such as *respected*- -. This is equivalent in scalar terms to *disrespected*++, yet it does more than convey disrespect; it also captures a failure to be as respected as one expects. An actor can now make an interpretative choice to show *respected*- - with a gesture that instead implies *respected*++, while relying on narration and dialogue to make the disrespect clear.

A metaphorical response is appropriate when a character's inertial valence is greater (by Δ or more) than the specific valence of the current action, and there is a need to dramatize this lagging emotional load. This dramatization is achieved by enacting a different but similar action with a valence (for that character) that is closer to its inertial valence in context. For instance, at plot point #10 in Fig. 2, an actor might react with gestures that suggest a character is despised (*disrespected*+++) rather than merely resented (*disrespected*++).

We estimate the opportunities for the performers to make interpretative choices by analyzing 10,000 generated stories with a setting of $\Delta = 2$ (so a character's inertial valence

must shift by 2 points or more to enable irony or metaphor). In this sample of 10,000 stories, 47,532 interpretative choice points are identified: 34,810 support metaphor, and 22,352 support irony. At this setting, a story has 4 to 5 opportunities for interpretative choice; 3 to 4 for metaphor and 1 or 2 for irony. We can adjust the Δ parameter to allow for more or less interpretation by the actors. Thus, when $\Delta = 1$ the average story provides 10 opportunities for interpretative choice, 8 to 9 for metaphor and 3 to 4 for irony (the two are not mutually exclusive). When $\Delta = 3$, the average story offers just 1 or 2 opportunities for interpretative departures from the script. The ideal setting for Δ, i.e. the setting at which audiences are happiest with the performers' use of metaphor and irony, has yet to be experimentally validated in user studies.

Evaluating Action At A Distance

Most gestures are ephemeral and transient, performed once and quickly forgotten. Few, if any, are persistent from one action to the next. While a performer can rely on gestures to express the emotions of the moment, they do not capture the inertia of the story so far. Actions that do not persist are not summative, and so do not reflect an inertial view. Ideally, if gestures are to be chosen on the basis of an interpretation of story events, some should also reflect this interpretation.

However, there is a class of physical actions that reflects a summative view of a story and its character interrelations at a given time: spatial movement to and fro on a stage. We noted earlier that some actions evoke emotions that move characters closer together or further apart in abstract terms, and this metaphorical movement is easily translated into the relative physical movement of robot performers. Motion of this kind is persistent: if a performer steps forward or back on stage, it will hold this position until it later moves again. Motion of this kind is also relative, and thus interpretative: as each performer moves back or forth to reflect their emotional reactions to an action, their relative position sums up the plot so far, and the emotional inertia that shaped it.

The choice to move a step forward or backwards is made by considering the emotional inertia a character brings to an action. Relative movement on the four emotionally charged scales is averaged to estimate an overall shift. If a positive shift exceeds Δ, the performer takes a step forward; if a negative shift exceeds $-\Delta$ they take a step backwards. Performers move as their current interpretations dictate, while leaving a gap that offers a more global perspective. This is how space is used coherently to convey meaning and emotion. Were the robots to invert this logic, their movements would lack schematic coherence, and we would expect an audience to be markedly less receptive to their overall performance.

This hypothesis has been validated using crowd-sourced user studies (Wicke and Veale 2021). In the *incoherent* condition, the robot performers do the opposite of what their spatial logic dictates. Conversely, the *coherent* condition has them follow this logic to the letter. We also conduct an evaluation in which it is a robot's gestures that are chosen coherently (to suit the action) or incoherently (at random). In each condition of each evaluation, judges are shown recorded fragments of a performance of the same Scéalextric story.

Methods The evaluations are crowd-sourced using AMT, *Amazon's Mechanical Turk*. Short 1-minute videos of robot performances, focusing on just two story actions, are shown to participants. Our pilot studies show that this format emphasises the movements of the robots and keeps participants engaged. A pool of $N = 160$ volunteers is evenly divided into four conditions: *coherent space*, *incoherent space*, *coherent gesture* and *incoherent gesture*. Each rater answers 14 questions about a video relating to just one condition. The first 7 measure the perceived attractiveness of the performance overall, while the second 7 relate to aspects of the embodied performance (e.g. whether the robots appear natural, or whether the participant would like to see the whole story). In addition, each questionnaire contains extra gold standard questions to weed out those who do not engage.

Analysis We excluded 42 participants for failing the gold-standard tests. The 118 valid responses are distributed across the four conditions as follows: *coherent space* (32), *incoherent space* (29), *coherent gesture* (29) and *incoherent gesture* (28). When each 14-item questionnaire is aggregated to yield a single appreciation score, a two-way ANOVA for *coherence versus incoherence* and *space versus gestures* shows a significant preference for the coherent performances (mean squares = 48.138, F values = 16.147 and p value <0.001). In particular, a significant difference is observed between the *coherent space* and *incoherent space* conditions (Bonferroni corrected p value = 0.047). Moreover, a post-hoc t-test shows significant differences between the coherent and incoherent conditions, while the effect size of Cohen's D = 0.197 indicates a small to medium effect that favours the coherent conditions. Conversely, since no significant differences between the spatial and gestural conditions are observed, it appears that audiences appreciate one as much as the other.

Results The results indicate that audiences appreciate performances that deliberately make interpretative use of space. Moreover, this schematic and logically simple use of space is just as effective as the use of a great many ad-hoc gestures.

Concluding Remarks

Some key distinctions in computational creativity invite a binary perspective when much greater nuance and gradation are called for. Consider, for instance, Boden's distinction between P- and H-Creativity (Boden 1990), which separates innovations that are original in a historical sense from those that merely seem novel to their producer. If novelty is instead judged on a graded scale, artifacts can be seen as more or less H- or P-Creative, with many exhibiting an affinity to both poles at once. *Mere generation* is yet another distinction that invites a binary perspective, and one that also reveals itself as gradated on closer inspection (Ventura 2016).

Indeed, applying as it does to a process as a whole, the distinction appears so binary and so judgmental that it no longer seems fit for purpose. As a replacement, we instead propose purely distinctive choice (or *pure disjunction*), since this affords greater nuance and greater scope for gradation. A process or a system can employ pure disjunction for one decision and interpretative choice for another, and so should be judged on how it achieves a balance of these alternatives.

Pure disjunction remains an attractive option for many generative tasks, and especially so for systems that aim to surprise with meagre resources. Unless interpretative choice is designed to surprise, its natural tendency is to tame the wildest excesses of random selection. Interpretation makes concessions to what it is familiar and normative, so as to couch the novel in the expected and achieve an "optimal innovation" (Giora et al. 2004). If too many concessions are made – as e.g., regarding gender roles in stories – interpretation bolsters the status quo. By paying little regard to what is normative in a relationship, a family or a workplace, pure disjunction can shatter norms to yield transgressive results.

As we must carefully balance interpretative and disjunctive choice to achieve the desired mix of novelty and value, we have presented a measure of *inertial valence* to dictate when a generative system should make the shift from one to the other. This measure is, in a sense, an "objective correlative" as defined by the poet T.S. Eliot (Barry 2002), allowing an actor's most dramatic choices to be adequately rooted in their character's unfolding relation to the underlying text. As such, it is instructive to note that non-computational practitioners, such as poets and critics, have also defined objective functions of their own, and computational creativity is a logical, computational extension of those critical approaches.

References

Barry, P. 2002. *Beginning Theory: An Introduction to Literary and Cultural Theory (2nd. Edition)*. New York: Manchester University Press.

Boden, M. 1990. *The Creative Mind: Myths and Mechanisms (second edition)*. Routledge.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Chen, M. X.; Lee, B. N.; and Bansal, G. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2287–2295.

Cook, W. W. 1928. *PLOTTO: the master book of all plots (2011 reprint)*. Tin House Books.

Csapo, A.; Gilmartin, E.; Grizou, J.; Han, J.; Meena, R.; Anastasiou, D.; Jokinen, K.; and Wilcock, G. 2012. Multimodal conversational interaction with a humanoid robot. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, 667–672. IEEE.

Delatorre, P.; Arfe, B.; Gervás, P.; Palomo Duarte, M.; et al. 2016. A component-based architecture for suspense modelling. In *Proceedings of AISB 2016's Third International Symposium on Computational Creativity (CC2016)*.

Fauconnier, G., and Turner, M. 2002. *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Garmendia, J. 2018. *Irony*. Cambridge University Press.

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.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, 117–124. Springer.

Häring, M.; Bee, N.; and André, E. 2011. Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *Ro-Man, 2011 IEEE*, 204–209. IEEE.

Kensinger, E., and Schacter, D. 2006. Processing emotional pictures and words: effect of valence and arousal. *Cogn. Affect. Behav. Neurosci.* 6(2):110–126.

McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Núñez, R. E., and Sweetser, E. 2006. With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive science* 30(3):401–450.

Oring, E. 2011. Parsing the joke: The general theory of verbal humor and appropriate incongruity. *Humor: International Journal of Humor Research* 24(2).

Ortony, A.; Clore, G.; and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Veale, T.; Wicke, P.; and Mildner, T. 2019. Duets ex machina: On the performative aspects of "double acts" in computational creativity. In *Proceedings of the 10th international conference on computational creativity*, 57–64.

Veale, T. 2012. *Exploding the creativity myth: The computational foundations of linguistic creativity*. Bloomsbury.

Veale, T. 2017. Déja vu all over again. In *Proceedings of the International Conference on Computational Creativity*.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24. Sony CSL, Paris.

Wicke, P., and Veale, T. 2018. Interview with the robot: Question-guided collaboration in a storytelling system. In *ICCC*, 56–63.

Wicke, P., and Veale, T. 2020. The show must go on: On the use of embodiment, space and gesture in computational storytelling. *New Generation Computing* 1–28.

Wicke, P., and Veale, T. 2021. Creative action at a distance: A conceptual framework for embodied performance with robotic actors. *Frontiers in Robotics and AI* 8:115–136.

Wilcock, G., and Jokinen, K. 2013. Wikitalk human-robot interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 73–74. ACM.