Advances in Cognitive Systems 4 (2016) 57-74 Submitted 8/2015; published 6/2016

Humor: A Dynamic and Dual-Process Theory with Computational Considerations

Boyang Li ALBERT.LI@DISNEYRESEARCH.COM
Disney Research, 4720 Forbes Avenue, Pittsburgh, PA 15213

Abstract

The cognitive mechanism of humor has been studied for centuries, with multiple seemingly incompatible theories proposed. However, none of existing theories is capable of explaining all empirical evidence. Recent research suggests emotional appraisals are tightly coupled and closely interact with other types of cognitive processes to create complex emotions and affects. This entangled nature contributes to the difficulty of humor research. In this paper, I attempt to provide a single, unified framework of humor, grounded in recent theoretical developments on emotion and dual-process cognition. I propose that humor comprehension consists of a quick succession of four major stages: surprise, reflection, dismissal, and compensation. The theory provides a modern update on existing theories of humor and is capable of explaining several phenomena that cannot be easily explained by existing theories. Finally, I outline a computational system that recognizes humor in the form of puns.

1. Introduction

Humor represents one of the most fluid and creative aspects of human intelligence. Popular fiction often portrays artificial intelligences as capable but humorless. As such, the study of humor may open a door to understanding the mechanism and organization of human cognition, as well as facilitate its replication. Theories of humor date back at least to Plato's *Philebus* and Aristotle's *Poetics*, both promoting the superiority theory (e.g., Bain, 1875), which posits that we laugh at the misfortune of other people. Since then, an abundance of theories have been proposed, ranging from the release of psychic or nervous energy (Spencer, 1860; Freud, 1928; Berlyne, 1972) to the formation of an incongruity that is later resolved (e.g., Koestler, 1964; Suls, 1972; Minsky, 1984; Veatch, 1998). Each theory seems to possess some explanatory power, yet none can satisfactorily encompass all empirical evidence and provide a unified explanation. Reviewing these theories, one is reminded of the ancient fable of four blind men and the elephant.[1]

Several new theories have been proposed in the past few years (McGraw & Warren, 2010; Hurley et al., 2013; Topolinski, 2014), attempting to provide a more unified explanation for humor. A common issue with these theories is that humor is still treated as an independent affect, created by an independent cognitive subsystem that serves an independent function. In contrast, recent research suggests that emotional appraisal works closely with other cognitive subsystems to create

1. Four blind men tried to figure out the shape of an elephant. They respectively touched its ear, nose, body, and leg, and all claimed an elephant is just like the body part they felt.

a rich emotional experience (Barrett, 2011; Scherer, 2001; Marsella & Gratch, 2009; Cunningham et al., 2013). In this paper, I provide the first attempt to explain humor as the result of interaction between widely recognized cognitive subsystems. My aim is to develop a more parsimonious theory of humor, which also sheds light on the development of AI systems that can create and understand humor.

The theory in this paper is built on top of two main theoretical foundations. The first is the dual-process theory (Stanovich & West, 2000; Evans, 2003; Kahneman, 2011; Evans & Stanovich, 2013), which states human cognition contains a set of automatic, effortless, fast, and intuitive processes, and a set of deliberate, effort-hungry, slow, and rational processes. The second is theories on emotional dynamics, including their construction from interactions among primitive processes (Barrett, 2011; Scherer, 2001; Marsella & Gratch, 2009; Cunningham et al., 2013).

My main argument is that the comprehension of humor is a dynamic process. It starts with a surprise that is sufficient to confuse the automatic processes and engage the deliberate processes, followed by a quick realization that the surprise is not worthy of further mental effort, which disengages the deliberate processes and produces an amplified positive emotion. This is the basic form of humor, with variations being produced by different realizations. The surprising stimulus may be trivial because we discover a logical flaw, malicious intent, stupidity, or social inappropriateness. These realizations further compound and enhance the effects of humor.

In this theory, humor comprehension employs a sequence of emotion appraisals and other cognitive inferences, which work together to produce a quick succession of emotions. To my best knowledge, this is the first attempt at explaining humor in terms of recent cognitive science theories and as interactions between multiple cognitive functions. We need not theorize a special standalone cognitive process for humor. Besides its parsimony, this theory also gains explanatory power by subsuming many existing accounts, including the superiority theory, the release theory, and the incongruity-resolution theory. It is also capable of explaining phenomena difficult to fit into existing theories, such as frustration smiles and humor's persistent appeal after repetition.

In the next section, I review relevant research in human cognition. After that, I introduce my theory of humor and compare it to existing accounts, discussing evidence that supports it. Finally, I discuss the implications for the study of congitive systems that aim to reproduce humor-related abilities in humans and outline a computational system for detecting puns. The theory is compatible with major cognitive architectures (Langley et al., 2009; Cox et al., 2011), although it has not been implemented in one yet.

2. Theoretical Background

In this section, I introduce the two main theoretical foundations of my theory on humor: the dual-process theory, and theories on emotion dynamics and emotion construction.

2.1 Two Types of Cognitive Processes

The dual-process theory (Stanovich & West, 2000; Stanovich, 2011; Evans & Stanovich, 2013; Kahneman, 2011) can be summarized as the co-existence of two types of processes in human cognition. The two types are often referred to as implicit and explicit or as automatic and deliberate. In this

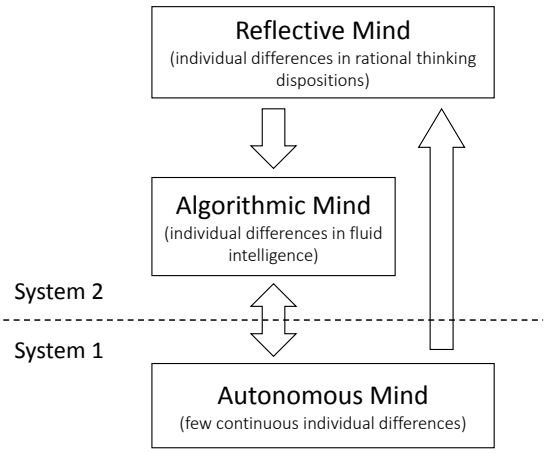

Figure 1. A tripartite model of the two-process theory. Adapted from Evans and Stanovich (2013).

paper, I refer to them as Sytem 1 and System 2. The former requires little mental effort and attention, works automatically, and responds quickly to external stimuli; it is relatively inflexible and error-prone when dealing with unfamiliar problems. The latter requires substantial mental effort; it works slowly and sequentially, but is flexible enough to handle novel and complex problems. System 1 does not utilize working memory, whereas System 2 requires it (Evans & Stanovich, 2013). As System 2 is tiring, we are inclined to delegate tasks to System 1 when possible.

The dual-process theory was proposed to explain the individual differences in solving cognitive puzzles designed to induce erroneous judgments. For example, the famous "Linda test" (Tversky & Kahneman, 1983) describes a woman with features stereotypically associated with a feminist. Participants are asked whether she is more likely to be both a feminist and a bank teller or just a bank teller. Although the second option cannot be less probable than the first option, most participants choose the first option. Some participants, however, can find the correct answer. The dual-process theory posits that individual differences mainly exist in System 2. Figure 1 shows a tripartite model proposed by Stanovich (2011). System 1 is shown as "autonomous mind" where individual differences are small. Individual differences, such as working memory and tendency to think extensively and gather evidence, are included in the "algorithmic mind" and the "reflective mind" . Those differences are the reason that some people can find the correct answer.

Several work studied the interaction between the dual processes and emotions. For example, positive emotions induce more automatic processing (i.e., System 1), whereas negative emotions induce deliberate processing (i.e., System 2), possibly because a negative emotion signals insufficient understanding and the need for more elaborate processing (Hullett, 2005). Surprise is believed to provide an important signal that triggers System 2 (Lieberman, 2003). When an expectation is violated, this indicates a novel situation that the associative System 1 cannot handle. For instance, when we try to open a door by turning a door knob, the action is automatic as long as the knob works. However, if turning the knob does not open the door, then deliberate processing by System 2 is activated. The engagement of System 2 by surprise is a key element in my theory on humor.

2.2 Theories on the Dynamics of Emotions

The recent constructionist theory of emotion (Barrett, 2006, 2011; Cunningham et al., 2013) claims that emotions are not atomic, indivisible, and categorical entities. Rather, emotions are constructed by the interplay of primitive cognitive processes, many of which are not dedicated to emotions. Physiological responses and brain scans for supposedly the same emotion can differ significantly depending on interactions between different processes (Barrett, 2006). However, so far this theory has not provided an exhaustive list of processes whose interactions produce complex emotions.

On the other hand, appraisal theories provide a list of appraisals, or cognitive evaluations and estimates responsible for creating emotions. In line with the constructionist theory, the EMA model (Marsella & Gratch, 2009) treats emotions as the results of continuous interactions between a set of basic emotional appraisals and other complex cognitive processes. The appraisal mechanisms check a number of variables that are important for the formation of emotions, taking inputs from the external stimuli, bodily responses, and results of other cognitive processes. The appraisals are fast compared to other cognitive processes and can be executed as soon as other processes produce results. Emotions are created and processed in an iterated cycle of appraisal, coping, and reappraisal. The appraisals in the EMA model include relevance, valence, intensity, future implications, blame/responsibility, and power/coping potential. The component process model (CPM) (Scherer, 2001, 2009) investigates the relative speed of different appraisals. It contains four appraisal checks happening in an invariant order from fast to slow, namely: novelty and relevance, goal conduciveness, coping potentials and power, and adherence to personal and social standards. Although the checks happen in a fixed order, they constantly interact with other processes and may repeat. Wessel et al. (2012) found novelty to be similar to expectation violation and surprise, as the three share neural circuits. In the Iterative Reprocessing model (Cunningham et al., 2013), information is processed in cycles to create emotions, so that the boundary between emotional processes and nonemotional processes is blurred.

Unifying the dual-process theory and the emotion theories is beyond the scope of this paper. My theory on humor critically relies on these following propositions from the theories:

1. Emotional appraisals interact with other cognitive processes to produce subjective emotional experiences. The appraisals evaluate results of other processes as they become available.
2. Emotion appraisals exhibit temporal regularity. The appraisal for expectation violation or surprise and the appraisal for relevance of an event are among the fastest. The two appraisals execute without much conscious effort and belong to System 1. The appraisal for personal and social norms is slower than the two.
3. Many other cognitive processes, especially System 2, are slower than the surprise appraisal, so they can produce results only at later time.
4. Our minds do not utlize deliberate processing from System 2 all the time, but it can be triggered by surprise.

These findings by modern cognitive science and neuroscience form a foundation for a new theory on humor, as presented below.

3. A Synthetic Cognitive Theory of Humor

I propose that humor and the associated affect, mirth, is not an independent emotional or affective category. Instead, mirth is the result of a quick succession of several emotion appraisals and cognitive processes. In this section, I introduce a synthetic cognitive theory of humor. The following pun can serve as an example to illustrate the theory:

Example 3.1. I asked if I was a gifted child, and Dad said we wouldn't have paid for you. (Vaid et al., 2003)

The first emotion in the sequence is surprise, arising autonomously from System 1. The importance of surprise in humor comprehension has been noted by many humor theorists (Minsky, 1984; Veatch, 1998; Huron, 2008; McGraw & Warren, 2010; Hurley et al., 2013). Surprise is a slightly negative emotion, indicating errors in the expectation formed by System 1 (Holroyd & Coles, 2002). Neuroscience evidence suggests brain regions responsible for surprise play a role in humor comprehension. Error detection, novelty and surprise are mainly processed in the anterior cingulate cortex (ACC) (Holroyd & Coles, 2002; Wessel et al., 2012). Some brain imaging studies (Mobbs et al., 2003; Watson et al., 2007) also found ACC activities during reading of humorous materials.

Reading Example 3.1, it is clear that the surprise happens at the end of the sentence. As we read the phrase "gifted child", our associative reflex leads us to one particular meaning of the word "gifted". However, given our interpretation, the phrase "paid for you" at the end of the sentence does not make sense. This probably causes many readers to stop and think, or even go back and read the sentence again, more carefully this time.

Surprises may indicate cognitive errors or physical threats for which we are unprepared for. The fast appraisal of surprise initiates System 2, which attempts to deal with this unfamiliar situation (Lieberman, 2003). System 2 strains the mind; it makes use of working memory and is correlated with dilated pupils, increased heart rate, and higher consumption of glucose (Gailliot & Baumeister, 2007; Kahneman, 2011). We are under stress to find out the source of the error and correct it, in the hope that we can learn and improve performance next time.

Although System 2 is usually slow, the joke is easy to make sense of. After a short period of reflection, System 2 realizes that the surprising stimulus is bogus: there are really no cognitive errors to correct and no threats. In Example 3.1, we attribute the surprise signal the two meanings of the word "gifted", which we already know, or to the intention of someone who played the pun on us. We probably will not find it funny if we do not know both meanings, say because English is not our native language. In addition, the father's behavior is also a little malicious, and we realize it should not be emulated for purposes other than joking. In other jokes, we can often attribute the error to the stupidity and social inferiority of story characters, so we dismiss any opportunities to change our understanding of the world or cope with an external threat. Therefore, we appraise the situation as irrelevant and not worth of further attention, and System 2 is deactivated.

A check for whether a stimuli is relevant for further processing exists in both the EMA model (Marsella & Gratch, 2009) and the CPM model (Scherer, 2001, 2009). In the CPM model, the relevance appraisal is a fast appraisal happening earlier than many other appraisals. However, the appraisal is based on the correct meaning of the sentence as inferred by the slow System 2. That is why the dismissal is the third stage of humor comprehension.

In some cases, we cannot dismiss the surprising stimulus as irrelevant, which in turn negatively influences the perception of humor. Proulx et al. (2010) find that, after reading a parody by Monty Python, those who perceive less threat to their values (and hence less need to reaffirm them) find the story to be funnier. The reason, according to the present theory, is that those who perceive threats to their values cannot dismiss the surprise as nonserious and irrelevant. The surprise-dismissal mechanism is similar to the incongruity-resolution theory discussed in Section 4.2. In comparison, my theory predicts that humor does not arise if the surprise is satisfactorily resolved and understood but not dismissed. Some surprises, like those in puzzles and riddles, do lead to genuine opportunities of learning:

Example 3.2. What walks on four legs in the morning, two legs in the afternoon, three legs in the evening, and no legs at night?

People unaware of the riddle's answer are often surprised. After hearing the answer, the surprise is resolved, but the riddle is still not funny because people realize that the surprise served its purpose and that they learned something from the answer.

But why is the quick dismissal and disengagement of System 2 processing funny? This can be understood in terms of the trampoline effect, proposed by Huron (2008). As discussed earlier, System 2 consumes mental effort and strains the mind. Realizing there was no cognitive error disengages System 2 and leads to relief, a positive emotion. As surprise is a slightly negative emotion, when the system adjusts from negative to neutral or slightly positive, it can overcompensate and reinforce the subsequent outburst of positive emotion. This is the trampoline effect. Topolinski (2014) studies the fluency of humor processing and shows faster realization of a joke's meaning leads to higher funniness rating. My theory explains this finding by assuming that the negative emotion, resulted from both surprise and the tiring cognitive load, accumulates as System 2 looks for the answer. Without a quick resolution, the accumulated negative emotion cancels out the positive emotion resulted from the joke.

Returning to Example 3.1, a significant portion of the mirth derives from social inappropriateness, in that the father implies his child is not worth much, and the self-deprecation of the speaker. A proponent of superiority theory would argue this is the entirety of the joke. However, this stance is incompatible with other theories and cannot explain all types of jokes (See discussion in Section 4.1). Feeling superior by itself, as occurs when beating an opponent in a difficult chess game, does not directly translate into humor.

I contend that the feeling of superiority, while cannot constitute the whole of humor, does compound the effects of humor. There are two possible ways to unify the superiority theory with the current cognitive theory: First, the social inappropriateness and implied low social status is a reason for dismissing the surprising stimulus. Therefore, it is part of the surprise-reflection-dismissal-compensation process rather than an independent function. Second, a well-timed judgment of superiority may be amplified by the trampoline effect (Huron, 2008). Mere superiority may not be funny, but the trampoline effect can make it funny.

One piece of neuroscience evidence comes from Moran et al. (2004). Based on fMRI imaging, they found humor understanding has two consecutive phases. In the first phase, humor detection, posterior temporal lobe and inferior frontal regions are activated. They theorized that the former is

responsible for retrieving expectation and the latter responsible for resolving ambiguities, which is in line with the incongruity-resolution account. Interestingly, the second phase, humor appreciation, activates the insular cortex, which is typically associated with pain perception and disgust rather than happiness. For this Moran et al. (2004) did not present a conclusive explanation. It is possible the insular cortex is activated by the detection of someone else's pain or our disgust toward someone else's behavior, which creates a superior feeling. In this view, this imaging result lends support to the relatively downstream position of superiority compared to surprise and reflection.

We can summarize the four stages of humor comprehension as surprise, reflection, dismissal, and compensation. In the surprise stage, an expectation created by the associative System 1 is violated, leading to surprise and the engagement of System 2. In reflection, System 2 makes sense of the surprising stimulus. In dismissal, the surprising stimulus, which now makes sense, is appraised as not worthy of further processing and dismissed as irrelevant. In compensation, one readjusts back from a slightly negative state to a positive state, producing a trampoline effect, which can be compounded by other factors such as superiority or pleasure of discovery. This dynamic process combines emotional appraisals and cognitive inferences that are part of System 2 to create the affective state known as mirth.

4. Existing Theories of Humor

In this section, I compare my theory to and contrast it with existing theories of humor. As many such accounts of humor exist, and major theories tend to have more than one variant, I will limit myself to the most influential candidates.

4.1 Superiority Theory

Superiority theory is one of the oldest and most popular theories of humor. Plato and Aristotle were both supporters of this theory. It can be illustrated with a classic ethnic joke:

Example 4.1. How many Poles does it take to screw in a light bulb? Five. One to hold the light bulb, and four to turn the table he is standing on.[2] (Attardo & Raskin, 1991).

According to Attardo and Raskin (1991), this joke was popular when Polish immigrants were discriminated against in early American history. At different historic periods, the joke had many variants, poking fun at different people.

However, the superiority theory have difficulties explaining all kinds of jokes. The feeling of superiority by itself, such as beating an opponent in a difficult chess game, does not directly translate into humor. Consider two puns that do not have an obviously inferior individual and the superiority theory cannot explain well.

Example 4.2. Two goldfish were in their tank. One turns to the other and says, "You man the guns, I'll drive." (Hurley et al., 2013)

Example 4.3. Photons have mass? I didn't even know they were Catholic. (Hurley et al., 2013)

2. It must be noted that the author does not support the racist view of this joke.

Admittedly, one can find the inferior individual with close reading. One may argue that we feel superior than the illusionary goldfish or the person who did not understand the meaning of "mass". Nevertheless, I find it implausible that the human cognition would spend significant effort to find an inferior individual during online processing of a joke.

In the proposed theory, superiority compounds the effects of humor, but is not a defining feature. Without clearly exhibited inferiority, humor is still possible. In these puns, the dismissal of surprise happens as soon as when we understand the double meaning of the word and make sense of the entire passage. We do not postpone laughing until we have clearly identified an inferior individual.

4.2 Incongruity Resolution

The *incongruity-resolution theory* (e.g., Koestler, 1964; Suls, 1972) states that humor is created by the forming of an incongruity that is subsequently resolved. This theory is also very popular and has many variants. Suls proposes a two-stage process, starting with an expectation violation (i.e., an incongruity), followed by problem-solving activities that reconcile the incongruity. Minsky (1984) notes a frame shift in puns: we realize one meaning of the word is wrong and shift to another meaning, where each meaning is represented by a frame.

Example 4.4. Everyone had so much fun diving from the tree into the swimming pool, we decided to put in a little water. (Binsted et al., 2006)

In the example above, readers initially has an image of a swimming pool filled with water, as people for most of the time jump into water. As soon as they reach the end of the sentence, the readers realize the pool was empty, and switch to the frame of suicide. In another influential theory, Attardo and Raskin (1991) claim humor is created by the opposition of two scripts, and the two scripts must be activated at the same time in the reader's mind.

The contribution of this paper over existing incongruity-resolution theories is twofold: First, it clarifies the theory and grounds it in modern cognitive science findings. My theory requires the surprise to be appraised as trivial and not an opportunity for learning, whereas the incongruity-resolution theory only requires it to be "resolved". This helps to elucidate the theory, differentiate humor from riddles and puzzles, and explain the fact that jokes can be repeatedly funny (see Section 5.2). In Attardo and Raskin's (1991) theory, it is not clear how two opposing scripts can co-exist. With the dual-process theory, it becomes clear that System 1 and System 2 can produce different interpretations, and only one is active after the punchline (Giora, 1991). Second, this paper reconciles the incongruity-resolution theory with other competing theories, including the superiority theory, the release theory, and the recently proposed theory of mistaken revelation (Hurley et al., 2013).

Another variant of the theory has to do with morality and social norms. Veatch (1998) argues humor is the simultaneous recognition of a violation of moral norms and an interpretation that the situation is normal. As a recent update to Veatch, McGraw and Warren (2010) propose a two-step process for humor comprehension. First, a violation of our expectation or social norms happens. This violation threatens to challenge our moral standards, humiliate us or shift our own beliefs about the world, which are all undesirable. However, we soon realize this violation is benign. For example, the person who was ridiculed is psychologically distant from ourselves. This leads to the feeling of mirth.

My theory is inspired by McGraw and Warren (2010), but offers a more restricted definition for "benign violation". Our expectations are violated daily by events such as a bus running late or a computer being unresponsive. Most of these violations do not cause severe consequences, and could be called benign, but they are not funny. Mirth only happens when the surprise itself is recognized to be fake and irrelevant. Another difference comes from the speed of the emotional appraisals involved. In the CPM model, the appraisal of social norms is the last of the four appraisals. Empirical evidence (Scherer, 2001, 2009) suggests the cognitive process for determining if a behavior is approved by social norms is slower than processes associated with earlier checks. Therefore, I postulate this check usually comes later in humor comprehension. Its main effect is not to cause surprise or trigger System 2, but to compound the effect of the positive emotions produced by earlier processes.

4.3 Release and Relief

The *release theory* (Spencer, 1860; Freud, 1928) claims that humor is caused by release of wrongly mobilized psychic energy. The slightly different *relief theory* (Berlyne, 1972; Meyer, 2000) states that humor is caused by a release of nervous energy, and is employed in revealing suppressed desires such as sexual desires. In the same vein, Immanuel Kant claims:

> Laughter is an affectation arising from the sudden transformation of a strained expectation into nothing (Kant, 1790).

This type of theory of humor can appear overly philosophical and archaic, even bordering on mysticism. They certainly hint at some kind of mental effort, but descriptions such as "into nothing" and "psychic energy" seem too vague to be of any value. Nevertheless, when grounded in the dynamic and dual-process theory proposed in this paper, these claims can make sense.

As noted earlier, System 2 is engaged when surprised. The use of System 2 requires conscious effort. Thus, the utilization of System 2 can be described as "mobilizing psychic energy". Kant's assertion of "nothing" coincides with the dismissal stage, which stops System 2 from processing of the surprising stimulus further. It is therefore not unreasonable to describe the mobilized energy as "released". In this sense, my theory provides a refreshed, modern view of the relief and release theories of humor.

4.4 Revelation of Mistaken Beliefs

Minsky (1984) may have been the first to suggest humor helps us learn to censor ridiculous thoughts. Developing the idea further, Hurley et al. (2013) propose a theory of humor and its evolutionary origin. According to this theory, humor is created when we realize one of our beliefs, which entered our mental space without our awareness, is wrong. Evolution has made recognition of mistaken beliefs fun, so as to encourage it. Thus, the sense of humor is evolutionarily adaptive. To Hurley et al., the basic form of humor is that I look for my glasses and find them on my ears.

My theory is influenced by and shares some similarities with Hurley et al. Indeed, surprise and error are closely related and share the same neural circuitry (Wessel et al., 2012). However, we also differ in important ways. I will not argue that the detection of errors, especially our own errors, is itself funny. Numerous studies on phenomena of confirmation bias, such as attitude polarization

(Lord et al., 1979) and persistence of discredited beliefs (Ross & Anderson, 1982), show that people are biased against recognizing their own mistakes. If a mechanism has been evolved to encourage us to recognize our own mistakes, it is not working very effectively.

My theory posits that the violated expectation is created by System 1, which often make mistakes, and quickly corrected and dismissed by System 2. Hence, we do not react to this violation of expectation overly negatively. In other words, humor comprehension does involve the realization that System 1 made an incorrect expectation, but this realization alone is not sufficient for humor. Moreover, I do not argue humor is created by evolution as a separate mechanism for detecting error. It may be a side product of our ability to detect expectation violation or evolved for its social function. Section 5.3 discusses social functions of humor.

My theory suggests a different basic form of humor that is illustrated when you raise a baby high in the air and lower him/her quickly. The baby is initially surprised, even slightly scared, by being raised. However, he/she then realizes there is no real danger. The giggle of the baby is the purest form of humor, uncompounded by factors like superiority.

5. Interpreting Humor Phenomena

A theory must be tested by its consistency with empirical data. In this section, I will show that my theory of humor can explain several curious phenomena that are difficult to explain under previous frameworks. The most powerful evidence includes the recently reported frustration smiles and the fact that jokes are often still funny after repetition. Furthermore, this evidence shows that constituents of the theory, including the dual process and the dismissal stage, are necessary to explain the mechanism of humor.

5.1 Frustration Smiles

In a study that aimed to induce frustration, Hoque et al. (2012) put participants to solve impossible reCAPTCHAs, and recorded their facial expressions and actions with a webcam. After repeated failures at the reCAPTCHA, 90% of their participants produced smiles during the experiment, even though the self reports contain only strong frustration. The smiles are real, utilizing muscles around the eye and the mouth. One author noted that the cognitive mechanism underlying these "frustration smiles" has been difficult to explain (Picard, personal communication).

I believe frustration smile results from emotion regulation (Gross, 2007; Gyurak et al., 2011). When we experience a negative emotion, we are motivated to cope with it. One possible coping strategy is reappraisal, which actively modifies the results of emotional appraisals before the effects of emotion are fully appreciated (Gross, 2002). Instead of admitting that they failed at solving the reCAPTCHA, the participants in the frustration experiment tell themselves that this is apparently a bug in the program or a prank. The coping strategy of trivializing the stimulus serves the same function as the dismissal stage of the humor process. As a result, the participants experience genuine mirth and exhibit genuine smiles.

The creation of frustration smiles follows the same surprise-reflection-dismissal-compensation process. However, the dismissal derives not from a direct emotional appraisal but a reappraisal in emotional coping. This suggests there can be multiple mechanisms responsible for the creation of

humor as long as they follow the general pattern and produce the same output. This demonstrates the universality of the proposed four-stage pattern and provides strong support to the current theory.

5.2 Jokes are Repeatedly Funny

If, as claimed by Minsky (1984) and Hurley et al. (2013), humor is a mechanism for us to discover cognitive errors and learn not to make them again, the effect of humor should decrease as a joke is repeated. If humor exploits our expectation, then it seems we should not create the same expectation again, but several studies show that it is not always the case. Belch and Belch (1984) found that low to medium levels of repetition actually increase the evaluation of humorous commercials, but high levels of repetition decrease their ratings, while Zhang and Zinkhan (1991) found humor of TV commercials is unaffected by repetition. Recently, using facial recognition software, Picard (personal communication) and colleagues find the second viewing of TV commercials can often increase the perceived joy.

Minsky (1984) has hypothesized that some parts of our cognition may not learn very quickly. This paper provides a concrete working mechanism to that theory: As System 1 is associative and inflexible, it defaults to retrieve the most frequent solution, such as the most likely word sense based on the immediate context. If the audience of humor do not consciously suppress the automatic response of System 1, they will still experience surprise, which triggers the humor response. However, as the audience have seen the joke before, their System 2 can make sense of the surprising stimulus faster than when it was encountered the first time. This makes the second encounter funnier than the first due to the fluency effect (Topolinski, 2014). However, continued exposure will eventually train System 1 and the joke will be "worn out" (Belch & Belch, 1984).

5.3 Social Functions of Humor

Research shows that laughter and humor are significantly influenced by the social setting and serve social functions. Zhang and Zinkhan (1991) found jokes are funnier when someone else is present. Butcher and Whissell (1984) found the funniness rating of TV commercials increases with the number of viewers. Fridlund (1991) found that even the presence of an imagined friendly companion can potentiate the smile action, regardless of the strength of the self-reported emotion. Meyer (2000) suggested that humor can create social bonding among those who laugh together as well as alienation for the butt of the joke. Multiple researchers (Miller, 2000; Bressler et al., 2006; Li et al., 2009) found that humor play an important role in how humans select their mates.

However, the question why humor should play such roles in human communication and reproduction has been left unexplained. The present theory suggests an asymmetry between the cognitive loads for joke crafting and joke understanding. The understanding of the joke partly utilizes the autonomous and effortless System 1, and it utilizes System 2 only briefly. In contrast, the joke teller must walk a thin line. A good joke must be difficult enough so System 1 does not understand it, and easy enough so System 2, being slow in nature, does not take too much time to understand it. The joke should preferably utilize other appraisals, such as superiority, to increase its funniness. Therefore, joke telling is a complex skill that can indicate intelligence.

In addition, understanding a joke often involve many cognitive appraisals, including surprise, moral judgments, and social identities. A joke is a condensed unit of information and a quick test

for shared values and attitudes. Two people laughing at the same joke suggest their appraisals work similarly, therefore they share values and attitudes. Not laughing often indicates otherwise. That is why laughing at one's boss's joke is widely believed to provide some career benefits. The role of humor in spouse selection can be explained as we look for people who are intelligent and share our values. The benefits of humor in social life and sexual selection provide one explanation for its evolutionary origin.

5.4 Individual Differences in Humor Appreciation

Although not well explored in the scientific literature (but see Forabosco & Ruch, 1994), anecdotal evidence suggests that individual differences exist in humor comprehension. People favor different types of jokes. The drama educator George Pierce Baker (1920) provided a differentiation of the so-called high and low comedy: "High comedy in contrast to low comedy rests then on thoughtful appreciation contrasted with unthinking, spontaneous laughter" (p. 236). Low comedy includes slapstick, farce, and jokes involving body parts, whereas high comedy often employs epistemic discovery such as those in puns. The dual-process theory is well suited for explaining such individual differences. As System 2 bears heavily on working memory, individuals with lower working memory efficiency and capacity (Jarrold & Towse, 2006) may take more time to understand complex jokes and thus find them less funny due to the lack of processing fluency (Topolinski, 2014).

6. A Computational Proposal

The present theory provides an account on the interplay between emotional appraisals and other cognitive inferences in the comprehension of humor. It underscores the complexity of human affects such as humor, suspense, group solidarity and identification, and the importance of building integrative architectures in order to model them. Phenomena like frustration smiles can only be understood in terms of interactions among multiple constituents of a system. The human ability to monitor our own performance, detect errors, adapt, and self regulate is central to this humor theory. Cognitive architectures built around expectation violation (e.g., Cox et al., 2011) could be a good starting point for a computational implementation of this theory.

Although the four stages of humor comprehension form a general pattern, depending on the joke, the exact working mechanism at each stage can be very different. For example, in a pun, the surprise usually comes from the actual meaning of the word being different from the first meaning the reader commits to. Comprehending puns requires linguistic knowledge and word sense disambiguation. In Example 4.4, the surprise comes from the background knowledge that people usually do not jump into an empty swimming pool. In slapstick or farce, the surprise may simply come from a loud noise or an unusual walking posture. Similarly, in the reflection stage, different kinds of problem-solving techniques may be employed to solve different problems. As an example, consider a joke that requires application of computer science knowledge:

Example 6.1. There are 10 kinds of people, those who understand binary and those who do not.

The difficulty of implementing a general system for humor comprehension lies in (1) acquiring required background knowledge and apply them at the right time, (2) implementing a system with

Table 1. Propositional representation of the first half of Example 3.1.

```
person(I), person(dad), father_of(dad,I),
ask_if(speaker, has_attr(I, talented), dad),
interpreted_as(phrase("gifted child"), talented))
```

multiple cognitive functions collaborating to solve complex problems, and (3) modeling the correct temporal characteristics for the interactions among different functions. Instead of tackling general-purpose humor understanding, which is well beyond current capabilities of AI, in this section I outline a domain-specific system that focuses on puns, such as those reported by Vaid et al. (2003):

Example 3.1. I asked if I was a gifted child, and Dad said we wouldn't have paid for you.

Following the suspense understanding system Dramatis (O'Neill & Riedl, 2014), the humor comprehending system would read a joke in predicate logic. The literals would be read sequentially. The system would contain an online mode and an offline mode. In the online mode, which mimics the operation of System 1, a limited memory retrieval process would attempt to make sense of the current logical literal being read. A failure to make sense of any logical literal would trigger surprise, upon which the system would enter the offline mode. In this mode, the system would escalate to a heuristic search, which is much more elaborate, flexible and time consuming than the memory retrieval. This is intended to mimic the working of System 2 in the reflection stage. After the joke is understood, the system would enter the dismissal stage and look for reasons to dismiss the surprise. In the final stage, compensation, an overall humor rating would be computed.

Table 1 shows the representation for the first half of Example 3.1. As each proposition is read, a small semantic network similar to Story Intention Graphs (Elson, 2012) and Modified Event Indexing with Prediction (O'Neill & Riedl, 2014) would be constructed. This network represents the current understanding of the joke and is thus called an interpretation network, which is to be distinguished from a much larger semantic network representing the entirety of the system's knowledge. Figure 2 shows an example network constructed with the literals in Table 1. The network contains two types of nodes: *entity nodes*, such as `I` and `dad`, and *predicate nodes*, such as `AskIf`. The network also contains *intentional frames*, such as the one around `Agree` and `FeelGood`. An intentional frame contains several actions performed by a character to achieve a goal, which is represented by a logical literal (Riedl & Young, 2010). Edges between entity nodes denote relations between entities, such as `IsA`. Edges labeled with `#` denote positions of arguments with respect to a predicate. For example, the `HasAttr` predicate has two arguments: the first argument `I` and the second argument `talented`. An intentional frame can also be an argument.

In online mode, the system would make sense of input literals by retrieving elements, including nodes and edges, from the background knowledge network and inserting them into the interpretation network. Nodes and edges that are entailed but not explicitly represented in the input literals may also be inserted. As the elements are selected from the background knowledge network, they can spread activation back into that network. Only a small number of elements that receive the highest activation in the background knowledge network can be retrieved. This is to simulate the limited,

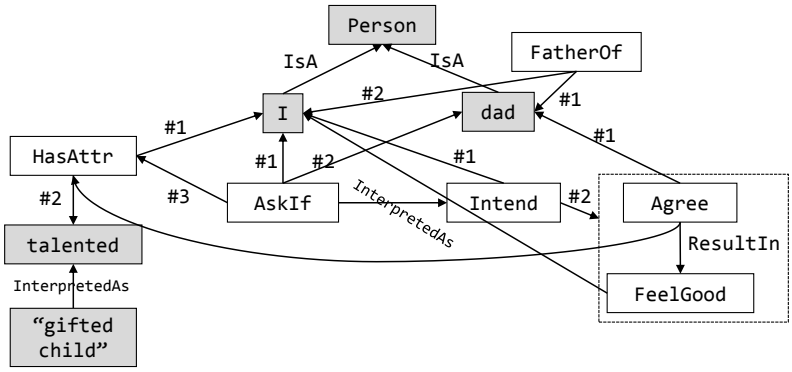

Figure 2. The interpretation network, constructed from input propositions and elements retrieved from the knowledge base. The positions of arguments to predicates are denoted with #1,#2, etc. Entity nodes are shaded.

associative inference in System 1. Backtracking is not allowed. Online sense making would fail as soon as a new input predicate node cannot be connected to any other predicate nodes in the interpretation network. This triggers surprise and offline sense making, which mimics the engagement of System 2.

The offline mode performs a deliberate search heuristically guided by activation, but it would not be limited to the first few options with the highest activation scores. Offline sense making can perform operations not allowed in the online mode, including removing elements from the interpretation network, adding alternative interpretations, and backtracking. Success is defined as completing an interpretation network where all story characters have intentions to support all actions they perform and each predicate node is related to least one other predicate node. In Example 3.1, a key inference step would be adding `Dad`'s intention as responding to an alternative explanation of what `I` said, where the phrase "`gifted child`" is interpreted as the child being given for free.

In the next stage, dismissal, the system would inspect the difference between the new solution and the old, failed solution, and look for reasons to dismiss the surprise. For example, by checking characters' intentions, it may conclude that `Dad` is purposely misleading as he responds to a unlikely explanation of what `I` said (i.e., `I` was given to `Dad` for free), so the surprise is dismissed as irrelevant for learning. Finally, in the compensation stage, the total strength of the joke is calculated by weighing the time spent in the offline mode against the strength of dismissal. More steps spent in the search would lead to less mirth due to reduced fluency of processing. A strong reason for dismissal (e.g., high absurdity) would lead to strong mirth.

4. Conclusions

Humor has been studied for centuries, with many alternative accounts for its working mechanism. Built upon recent cognitive theories, especially dual-process cognition and emotional dynamics, this paper provides a detailed cognitive account for humor that unifies different theoretical positions and

explains various related phenomena. I describe humor comprehension as resulting from complex interactions between multiple cognitive functions over a short period of time.

According to the present theory, mirth is created by a quick succession of four emotional and cognitive stages: surprise, reflection, dismissal, and compensation. In the surprise stage, an expectation produced by the automatic System 1 is violated, leading to surprise and the engagement of the deliberate System 2. In reflection, System 2 makes sense of the surprising stimulus. In dismissal, the surprising stimulus is appraised as irrelevant and not worthy of further processing. In compensation, as the mind adjusts from a slightly negative state to a positive state, a trampoline effect occurs, creating an outburst of positive emotion, which can be compounded by other factors such as superiority or the pleasure of epistemic discovery. Converging evidence from brain imaging, facial recognition studies, the effect of repetition on humor, and frustration smiles supports the plausibility of this theory.

Nevertheless, the theory should be considered as a work in progress rather than as a complete account. Further substantiation and investigation is needed, especially on the compounding effect of superiority and the mechanisms of the trampoline effect. Important questions include whether recognition of superiority is a cause for dismissal and whether it can happen after dismissal. It would also be desirable to enumerate possible causes for dismissal and determine if they work similarly and in the same time frame.

The theory makes a number of falsifiable claims regarding the temporal sequence of humor understanding and dual-process cognition. Both behavioral experiments and brain imaging can clarify the temporal sequence. For example, the theory predicts that if we completely remove the surprise, to the extent that System 1 finds it predictable, no humor would result, even with the presence of a superior feeling. The theory contends that surprise happens earlier than the engagement of System 2, which happens earlier than superiority. This may be verified by, as examples, brain imaging and measurment of pupil sizes, which enlarge when System 2 becomes active (Kahneman, 2011).

The current theory highlights the importance of studying complex affective responses, such as humor and suspense, in the context of interactions between cognitive processes and subsystems. Building off recent advances of cognitive science, we may finally begin to crack the age-old mystery of humor.

Acknowledgements

I am grateful for discussions and suggestions from Stacy Marsella, Brian Magerko, Pat Langley, Rosalind Picard, Maarten Bos, Lew Lefton, and Pete Ludovice.

References

Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: Joke similarity and joke representation model. *Humor*, *4*, 293–347.

Bain, A. (1875). *The emotions and the will (3rd ed.)*. London: Longsman & Green.

Baker, G. P. (1920). *The development of Shakespeare as a dramatist*. New York: The Macmillan Company.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*, 28–58.

Barrett, L. F. (2011). Constructing emotion. *Psychological Topics*, *20*, 359–380.

Belch, G. E., & Belch, M. A. (1984). An investigation of the effects of repetition on cognitive and affective reactions to humorous and serious television commercials. *Proceedings of the Eleventh Conference on Advances in Consumer Research* (pp. 4–10). Provo, UT: Association for Consumer Research.

Berlyne, D. E. (1972). Humor and its kin. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor*, 43–60. New York: Academic Press.

Binsted, K., et al. (2006). Computational humor. *IEEE Intelligent Systems*, *21*, 59–69.

Bressler, E. R., Martin, R. A., & Balshine, S. (2006). Production and appreciation of humor as sexually selected traits. *Evolution and Human Behavior*, *27*, 121–130.

Butcher, J., & Whissell, C. (1984). Laughter as a function of audience size, sex of the audience, and segments of the short film 'Duck Soup'. *Perceptual and Motor Skills*, *59*, 949–950.

Cox, M. T., Oates, T., & Perlis, D. (2011). Toward an integrated metacognitive architecture. *Advances in Cognitive Systems: Papers from the 2011 AAAI Symposium* (pp. 74–81). Arlington, VA: AAAI Press.

Cunningham, W., Dunfield, K., & Stillman, P. E. (2013). Emotional states from affective dynamics. *Emotion Review*, *5*, 344–355.

Elson, D. (2012). *Modeling narrative discourse*. Doctoral Dissertation, Computer Science Department, Columbia University, New York.

Evans, J. S. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.

Evans, J. S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

Forabosco, G., & Ruch, W. (1994). Sensation seeking, social attitudes and humor appreciation in Italy. *Personality and Individual Differences*, *16*, 515–528.

Freud, S. (1928). Humor. *International Journal of Psycho-Analysis*, *9*, 1–6.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*, 229–240.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review*, *11*, 303–327.

Giora, R. (1991). On the cognitive aspects of the joke. *Journal of Pragmatics*, *16*, 465–485.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281–291.

Gross, J. J. (Ed.). (2007). *Handbook of emotion regulation*. New York: Guilford.

Gyurak, A., Gross, J. J., & Etkin, A. (2011). Explicit and implicit emotion regulation: A dual-process framework. *Cognition and Emotion*, *25*, 400–412.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

Hoque, M., McDuff, D., & Picard, R. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, *3*, 323–334.

Hullett, C. R. (2005). The impact of mood on persuasion: A meta-analysis. *Communication Research*, *32*, 423–442.

Hurley, M. M., Dennett, D. C., & Adams, R. B. (2013). *Inside jokes*. Cambridge, MA: MIT Press.

Huron, D. (2008). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Jarrold, C., & Towse, J. N. (2006). The physiology of laughter. *Neuroscience*, *139*, 39–50.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kant, I. (1790). *Critique of judgement*. New York: Hafner Publishing Co. Republished in 1951.

Koestler, A. (1964). *The act of creation*. New York: MacMillan.

Langley, P., Laird, J. E., & Rogers., S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Li, N. P., Griskevicius, V., Durante, K. M., Jonason, P. K., Pasisz, D. J., & Aumer, K. (2009). An evolutionary perspective on humor: Sexual selection or interest indication? *Personality and Social Psychology Bulletin*, *35*, 923–936.

Lieberman, M. D. (2003). Reflexive and reflective judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. D. Williams, & W. V. Hippel (Eds.), *Social judgments: Implicit and explicit processes*. New York: Cambridge University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Marsella, S., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, *10*, 70–90.

McGraw, A., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science*, *21*, 1141–1149.

Meyer, J. C. (2000). Humor as a double-edged sword: Four functions of humor in communication. *Communication Theory*, *10*, 310–331.

Miller, G. F. (2000). *The mating mind: How sexual choice shaped the evolution of human nature*. New York: Doubleday.

Minsky, M. (1984). Jokes and the logic of the cognitive unconscious. In L. Vaina & J. Hintikka (Eds.), *Cognitive constraints on communication: Representations and processes*, 175–200. Dordrecht, Netherlands: Springer.

Mobbs, D., Greicius, M. D., Abdel-Azim, E., Menon, V., & Reiss, A. L. (2003). Humor modulates the mesolimbic reward centers. *Neuron*, *40*, 1041–1048.

Moran, J. M., Wig, G. S., Adams, R. B., Janata, P., & Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *Neuroimage*, *21*, 1055–1060.

O'Neill, B., & Riedl, M. O. (2014). Dramatis: A computational model of suspense. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 944–950). Québec City, Québec: AAAI Press.

Proulx, T., Heine, S. J., & Vohs, K. D. (2010). When is the unfamiliar the uncanny? Meaning affirmative after exposure to absurdist literature, humor, and art. *Personality and Social Psychology Bulletin*, *36*, 817–829.

Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, *39*, 217–268.

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research*. Oxford, UK: Oxford University Press.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*, 1307–1351.

Spencer, H. (1860). The physiology of laughter. *Macmillan's Magazine*, *1*, 395–402.

Stanovich, K. (2011). *Rationality and the reflective mind*. New Haven, CT: Yale University Press.

Stanovich, K. E., & West, R. F. (2000). Individual difference in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences*, *23*, 645–726.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In P. E. Goldstein & J. H. McGhee (Eds.), *The psychology of humor*, 81–100. San Diego: Academic Press.

Topolinski, S. (2014). A processing fluency-account of funniness: Running gags and spoiling punchlines. *Cognition and Emotion*, *28*, 811–820.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 239–278.

Vaid, J., Hull, R., Heredia, R., Gerkens, D., & Martinez, F. (2003). Getting a joke: The time course of meaning activation in verbal humor. *Journal of Pragmatics*, *39*, 1431–1449.

Veatch, T. C. (1998). A theory of humor. *Humor*, *11*, 161–215.

Watson, K. K., Matthews, B. J., & Allman, J. M. (2007). Brain activation during sight gags and language-dependent humor. *Cerebral Cortex*, *17*, 314–324.

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *The Journal of Neuroscience*, *32*, 7528–7537.

Zhang, Y., & Zinkhan, G. M. (1991). Humor in television advertising: The effects of repetition and social setting. *Proceedings of the Eighteenth Conference on Advances in Consumer Research* (pp. 813–818). Provo, UT: Association for Consumer Research.