A robot's sense-making of fallacies and rhetorical tropes. Creating ontologies of what humans try to say

Hoorn, Johan F.; Tuinhof, Denice J.

published in
Cognitive Systems Research
2022

DOI (link to publisher)
10.1016/j.cogsys.2021.10.003

document version
Publisher's PDF, also known as Version of record

Cognitive Systems Research 72 (2022) 116–130

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

A robot's sense-making of fallacies and rhetorical tropes. Creating ontologies of what humans try to say

Johan F. Hoorn [a,*], Denice J. Tuinhof [b,c]

[a] The Hong Kong Polytechnic University, Hong Kong
[b] Vrije Universiteit Amsterdam, the Netherlands
[c] Universiteit van Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Social robots
Logical fallacies
Metaphors
Reference
Sense
Maxim of quality
Tableau reasoning
Epistemics of the virtual

ABSTRACT

In the design of user-friendly robots, human communication should be understood by the system beyond mere logics and literal meaning. Robot communication-design has long ignored the importance of communication and politeness rules that are 'forgiving' and 'suspending disbelief' and cannot handle the basically metaphorical way humans design their utterances. Through analysis of the psychological causes of illogical and non-literal statements, signal detection, fundamental attribution errors, and anthropomorphism, we developed a fail-safe protocol for fallacies and tropes that makes use of Frege's distinction between reference and sense, Beth's tableau analytics, Grice's maxim of quality, and epistemic considerations to have the robot politely make sense of a user's sometimes unintelligible demands.

1. Introduction

There is this cartoon by Randy Glasbergen,[1] saying "I want a computer that does what I want it to do, not what I tell it to do!" Computer users sometimes mean something different from what they put into the computer while the computer executes commands literally, much to the user's aggravation. What if we could teach a computer to check what the user means if it encounters a command that would go against the ontology ('knowledge base') it knows the user normally keeps? That would lay the foundation of a robot that does not blindly execute commands nor tell the user wrong but deals with mistakes, fallacies, and figurative language from a position of understanding what it is that its user tries to convey (Angleraud, Houbre, & Pieters, 2019) (cf. the 'cognitive turn in logics' advocated by Magnani, 2015).

In the current paper, we will discuss four types of utterances that are hard to process by a computer, two of which come from logical error, two from associative combination-making. The logical fallacies we discuss are *ex-consequentia* reasoning ("That robot is polite because it waves goodbye" A → B, B, → A) and *inverse error* ("If you're not a robot, then you don't run on electricity" A → B, ¬A, → ¬B).[2] The rhetorical tropes are metaphors (A is B). Metaphors follow from the misattribution of an exemplar to a category (e.g., that man is a beast). Related tropes such as simile (A is like B) follow the same principle. One type is founded on 'missing the signal.' For instance, 'My robot is human too' poses that this particular robot (a fictitious being) belongs to the category of real humans. The other type of metaphor is constructed from a 'false alarm.' For example, 'My husband acts like a robot' attributes a real person to a category of fictitious beings named robots. Yet, logical fallacies and metaphorical tropes all originate from the way people acquire their knowledge and the way they validate it (their 'epistemics'), which is the hard part for a robot to grasp.

In the first half of the paper, we attempt an account of *why* user mistakes in logics happen and *why* non-literal comparisons such as metaphors and similes occur. In Fig. 1, we offer an overview of our reasoning. In the second half, we propose a 'fail-safe protocol' that tells a robot how to deal with these type of complicated utterances. In brief, the protocol compares the ontologies of user and robot and when mismatches occur, the robot suspends disbelief and maintains that the user 'intends to be truthful.' Then it detects the type of fallacy and/or analyzes the metaphoric expression, attempting to reconstruct the intended meaning from the user's ontology. Note that the protocol is still far off to be implemented (also check the Supplementary Materials). We offer a formal grounding for guiding techniques that may facilitate human-robot communication even though those techniques are

* Corresponding author at: Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
E-mail address: csjfhoorn@comp.polyu.edu.hk (J.F. Hoorn).

[1] https://s-media-cache-ak0.pinimg.com/736x/95/16/d2/9516d292f976ef8c5b060d5499767903.jpg

[2] The symbol ¬means 'not.'

https://doi.org/10.1016/j.cogsys.2021.10.003
Received 28 August 2019; Received in revised form 3 December 2019; Accepted 14 October 2021
Available online 29 October 2021
1389-0417/

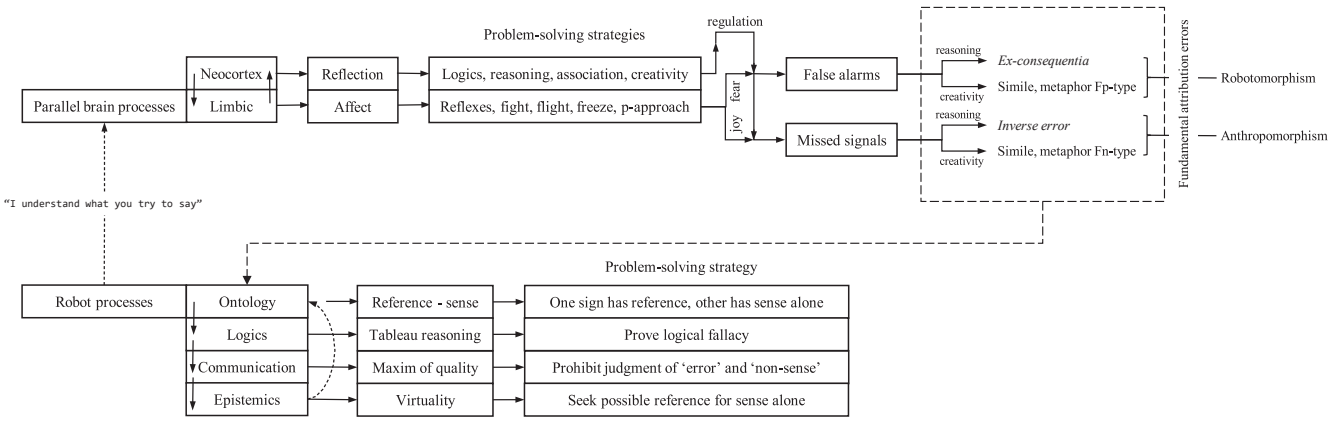


Fig. 1. Robot's sense-making of fallacies and rhetorical tropes.

underdeveloped still.

The account in Fig. 1 is a theoretical position, not a generally accepted fact. We will try to present evidence to support that position and if still wanting, we will indicate that the assertion is in need of corroboration. Our exploration of faulty and non-literal utterances starts with a discussion of two modes of information processing in the human brain, which we believe are responsible for producing fallacies and metaphors. This part has no application value but is the scientific backdrop that motivated the composition of the fail-safe protocol.

LeDoux (1996/1999) proposed two pathways of the emotional brain. The reflective mode relates to the neocortex and is more rational; the affective mode mainly concerns the limbic system and is more emotional. In Fig. 1, the brain is thought to run two modes of processing in parallel: reflective and affective (Konijn & Hoorn, 2017). At times, a network of 'valuing' systems (the neural systems that evaluate the affective side of information) shortcut the main circuitry for cognitive control (Crone & Dahl, 2012, p. 640). This may affect the way people perceive the 'reality status' of a media utterance or fictional character such as a robot (Crone & Konijn, 2018).

Both processes are always in function but one mode may take precedence over the other (Mujica-Parodi, Cha, & Gao, 2017). The limbic, affective process, solves problems that demand immediate action: fight, flight, freeze, or positive approach (e.g., to make up with someone). Its tactic is reflexive, taking things at face value (wysiwyg) and trading accuracy for speeded decisions (ibid.). The reflective, cortical system, controls the limbic system so that on second thought, mistakes are corrected or emotions are regulated (ibid.). The reflective mode is more concerned with logics and reasoning but also with association and creative problem solving (Pfenniger, 2001, p. 91).

Fig. 1 shows that when fear and joy arise from the affective process, the reflective process interferes to channel those emotions into the right direction (cf. Thompson, Uusberg, Gross, & Chakrabarti, 2019). The anger of a reasonable person should not lead to a fistfight but to a discussion. However, sometimes the affective process takes precedence and reflection may not be absent but is backgrounded at the least. At that point, mistakes easily happen (seeing too little, seeing too much) as accuracy is traded for rash action.

As an issue of changes in perceptual ability, for instance, emotional states such as anxiety are intolerant to uncertainty, biasing what people see and the way they see it (Cataldo & Cohen, 2015). Perceptually, anxious people seem to discriminate faces better than other stimuli such as houses (ibid.). Under stress, sensitivity may increase to discern emotional cues in other people's faces so to seek help or to detect a threat (Domes & Zimmer, 2019). In terms of signal-detection theory, the number of false alarms (i.e. seeing something that is not there) and of missed signals (i.e. overlooking something that *is* there) are likely to increase (Mujica-Parodi et al., 2017).

We argue that signal-detection faults may give rise to specific types of logical fallacies as well as to metaphors and similes, dependent on the problem-solving strategy of the reflective process. This is a bold statement and it is *not* said that signal-detection faults automatically give rise to fallacious and metaphoric utterances. We argue that false alarms in combination with reasoning activities likely produce *ex-consequentia* fallacies (A → B, B, → A). False alarms in combination with creativity lead to false-positive or *Fp-type* of metaphors (e.g., 'Humans are robots'). Missed signals combined with reasoning likely produce *inverse errors* (A → B, ¬A, → ¬B) and in combination with creativity to false-negative or *Fn-type* of metaphors (e.g., 'Robots are human'). These are new hypotheses that (therefore) have no direct empirical support. However, there is little use in measurement if a theory is not worked out yet, which is what we do here.

We also avow that the four types of (literally) false statements (Fig. 1, dashed box) may result into so-called fundamental attribution errors (Ross, 1977). People are inclined to 'naïve realism' (Ross, 2018), thinking their perceptions and the ensuing beliefs, preferences, priorities, and feelings are 'objective' (ibid.). For instance, people tend to think that someone's actions (e.g., late at a meeting) express what internally goes on inside that person (i.e. must be an indifferent person) and tend to overlook external explanations such as a person being forced by circumstances (e.g., a traffic jam). Reversely, people tend to attribute successes (e.g., win tennis match) to their own qualities whereas failure is blamed on the circumstances (e.g., a bad court) (Ross, 1977; 2018).

If we relate these findings back to robots, people make two types of attributions: They apply human-like qualities to a machine without considering that a mistake (e.g., Damiano & Dumouchel, 2018), disregarding that the robot may be operated by human hand, and missing out on the cues to being a robot (e.g., a synthetic voice). This is what is called *anthropomorphism* (e.g., Epley, Waytz, & Cacioppo, 2007).

People also identify in other humans their 'machine-like' qualities. Emotionally unavailable men are unequivocally typified as 'robot.' Those who base themselves on Western science mostly maintain a mechanized world view (see the discussion by Imanishi, 1941/2002, p. 7). In biology, medicine, (neuro)psychology, and in lay theory alike, the body and particularly the brain often are framed as 'machinery' that contains various 'mechanisms' (Hauser, Nesse, & Schwarz, 2017). Notions such as "brain mechanism" or "automated brain processes" are highly metaphoric uses of words that are strongly related to logical fallacies (see Hoffman, Cochran, & Nead, 1994, p. 186).

On occasion, humans robotize themselves so that the perceiver thinks there is a robot (false alarm) whereas in fact it is a human who does playacting. Wizard of Oz situations (e.g., Ishiguro's Erica), remote controlled robots (e.g., Nao/Zora) are all like a Mechanical Turk,[3] a

[3] http://www.bbc.com/news/av/magazine-21882456/meet-the-18th-century-chess-machine

chess-playing automaton built in 1770 by the Hungarian Kempelen Farkas for the Empress of Austria. The 'Turk' hid a human chess player inside and fooled people for about 84 years. People may feel deceived when humans pretend to be robots without telling: "…mere tricks; tricks inferior to many flights of hand…" (Thicknesse, 1784, p. 4).

While seeing more than meets the eye, these false alarms to being machines result into detecting robotic elements in other people, framing fellow humans as apparatuses in disregard of their organic origins ("My friend acts like a robot without any emotions").[4] This is what may be called *robotomorphism* (cf. Jiménez-López, 2009, p. 80), representing humans as stimulus–response machines (Von Bertalanffy, 1968, pp. 190-191), applying schemas and templates from machines and devices onto human conduct (cf. behaviorism).

How should a robot deal with utterances of its user that are literally and/or logically false (Fig. 1, dashed box)? We propose a four-pronged approach, that involves the way a robot updates its ontology about the user, the logics and communication rules it should apply, and the epistemic system that attributes 'believe' or 'empirical plausibility' to unintelligible user statements.

To do so, we work from Frege (1892) distinction between reference and sense. We apply Beth's (Beth, 1955) tableau analytics to prove a statement is false, yet keep the robot from telling its user wrong by introducing an intensional operator under Grice (1975) Maxim of Quality. We end with Hoorn (2012) knowledge theory that can make sense of fictitious, imaginative, and possible worlds to update the robot's ontology about its user again. Like this, we hope the robot might one day reply to its illogical user that speaks with rhetorical tropes: "I understand what you try to say."

2. Reflexive, reflective

The limbic system plays a pivotal role in the first assessment of the situation an organism is in. It specializes in unconscious, reflexive responses to react fast to, in particular, threatening situations (Mujica-Parodi et al., 2017). If a twig on the ground looks like a snake, it will respond to it as if it were a snake from a heuristic of 'better safe than sorry' (cf. 'negativity effect' in Ceschi, Costantini, Sartori, Weller, & Di Fabio, 2019). Because social interaction is crucial to survival for humans, recognition of facial expressions (angry, sad, happy) also goes through the limbic system first (Diano et al., 2017).

In their Experiment 1, Appel and Richter (2010) found that with high need for affect, people are transported easily into a story that is relevant to their beliefs so that the fictional narrative actually becomes influential for those believes. If people are lonely and the robot is framed as a companion, a lover, a grandchild, or something else that is relevant to the user's believes, people may easily go along with the doll play and their beliefs are impacted by their play act. Make-believe impacts beliefs about reality (Appel & Richter, 2010).

Konijn and Hoorn (2016, 2017) explain that emotions lend "realness" to the object of emotion such as a robot, signaling the user that something of "real" importance is going on. The observer interprets his or her personal physiological change as proof of reality and it influences the perception of the object that evokes the emotion (Crone & Konijn, 2018). When the affordances of a robot, the things you can do with it, are relevant to user objectives, the evoked emotions 'prove' to the limbic system that the robot has genuine traits ('It likes me!'), even if this is manifestly not so programmed.

However, this hardwired reflexive response does not stand on its own because it may be fast but it is not too accurate (Mujica-Parodi et al., 2017). The snake may be just a twig, the smile of the alpha male does not mean happiness but frustration, and a robot does not have feelings. Therefore, the reflexive and affect-oriented limbic system is kept in check by a more cognitive-reflective thinking mode (ibid.) to keep the affective assessment from making critical mistakes (cf. 'normative rationality' in Ceschi et al., 2019).

The neocortex has several problem-solving strategies in store that are more sophisticated than merely attack, embrace, flee, or sit still. It may rely on intelligence to reason logically from the givens to a conclusion: Snakes are not of wood. The twig consists of wood. Therefore, the twig is not a snake. Its solution may come from creativity when information is scarce and problems are underdetermined (Fig. 1): To ward off my enemies, I make a snake from this twig. Although affective and reflective processes run in parallel, sometimes the reflective mode is strongly suppressed (Mujica-Parodi et al., 2017) when something is acute (e.g., a car accident), when in fear (e.g., during a robbery), or in joy (e.g., when being love-struck).

3. Signal detection

Taking a twig for a snake is an example of a 'false alarm' or 'false positive' in signal-detection theory (Mujica-Parodi et al., 2017). Likewise, the statement 'my friend behaves like a robot' also is a false alarm, expressed in a simile. Under emotional circumstances, the signal-to-noise ratio deteriorates (Fig. 2). Signal in our case is the difference between a human and a robot, the noise being the variability in robots and humans showing resemblance (e.g., in appearance or intelligence).

For that part of the readership who sufficiently well understands signal-detection theory, the detailed exposition in the current section may be superfluous. Yet, it is potentially interesting for its application to the Turing Test and it explains the bias towards ambiguous stimuli as false positive or negatives in line with the neurological account of Mujica-Parodi et al. (2017).

Detection theory looks into the difference between the response distribution to something exceptional (the signal: a snake, robot) relative to the response distribution to everything normal (the noise: a twig, human beings). This difference often is expressed as the distance *d'* between the top of the two normal distributions of internal responses for, in the case of the human as robot, "Human" (the noise) and "Robot" (the signal plus noise). Distance *d'* is estimated with the standardized difference (*z*-scores) between the right-tail probabilities *(p)* on the normal distributions, where $d' = z_{(\text{false alarms})} - z_{(\text{hits})}$ and $z = (p - M) / SD$, assuming that $p \neq 1$ (hits only) and $p \neq 0$ (hits absent).

Index *d'* indicates personal sensitivity to the difference between, in our case, robots and humans (Fig. 2). Higher values for *d'* mean more sensitivity to the difference and less overlap between the distributions. There is better stimulus discrimination. Thus, the more signals (cues to a robot) amount on top of the noise (everything is human), the more the distribution for "Robot" moves to the right and distance *d'* increases (you see better it is a robot).

In a cognitive reflective mode, which is slower and more accurate, people are more critical to detail so that sensitivity to the difference is higher (*d'* increases). In reflective mode, people evaluate whether the signal–noise ratio is large enough to conclude that the difference (distance *d'*) is not due to chance. In their minds, hypothesis H1 states that robots differ from humans whereas H0 says that what you see is mere variability in humans and that you cannot tell that there is a robot around (cf. Turing Test). There should be enough cues to a robot to decide for "robot" and not for just another human being (perhaps it is someone with dyspraxia or an emotionally unavailable husband). In signal detection, the null hypothesis stands for true negatives: There is no signal; there is nothing at hand (Fig. 2).

In reflective mode, people maintain a conventional level of confidence that their observations are correct. They may say that 'It is not so that humans and robots do not differ (reject H0) but I might be wrong about it.' Put differently, they may say that 'Humans and robots do differ (accept H1) and I am for 95% sure of it' (so the probability they are wrong is about 5%, $p < .05$). If people cannot reject the H0, the agency is just another human being. The alternative explanation is that the

[4] https://www.quora.com/My-friend-acts-like-a-robot-without-any-emotions-is-it-normal

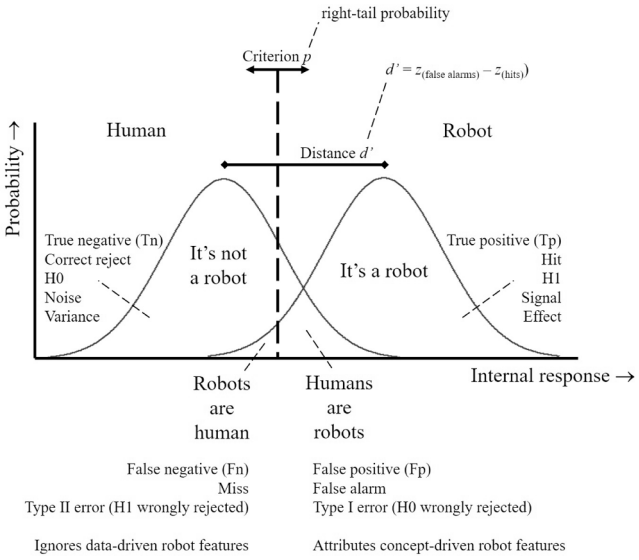


Fig. 2. Signal-detection account for observing a robot. Emotions make the distributions overlap more.

observer cannot prove that this one agency actually is a robot, for instance, because it has passed the Turing Test (Hoorn, Konijn, & Pontier, 2018).

In an affective reflexive mode, which is fast and imprecise, sensitivity *d'* decreases (Fig. 2). Distinctions are not that clear any more, the two distributions grow together, they 'blur,' and mistakes occur in the overlapping area.

Fredrickson and Branigan (2005) state that positive and negative emotions may widen or narrow someone's perceptions, respectively. Negative emotions narrow down the focus to the object of anger, fear, or anxiety, which is important for a fight, flight, or freeze response. Then again, negative emotions limit a person's worldview and the range of available thoughts and actions they can apply to the situation.

Positive emotions have the opposite effect. Fredrickson (1998, 2001) states that positive emotions make people 'look around' in their environment, broadening their worldview and expanding their thoughts and actions. People become more lenient in their cognitive associations and categorizations (Isen, Johnson, Mertz, & Robinson, 1985; Murray, Sujan, Hirt, & Sujan, 1990). In our case, the distinction between robot and human would not be that crisp any more. Positive emotions also make people more creative (Ziv, 1976, in: Fredrickson, 1998; Davis (2009)) compared to neutral or negative affect, which may give rise to more metaphor production.

When positive and negative emotions become intense, however, rational decision-making seems to become undermined (Bechara, 2004, 2005; Dolan, 2007; Dreisbach, 2006). People sometimes fear the introduction of robots for dehumanizing care or expected job loss. Although the data fearful people perceive may suggest otherwise, from a concept-driven perspective they see the influence of robots everywhere: People may fear that robots are taking over and that Artificial Intelligence (AI) will outsmart us all, that 'AI is our biggest existential threat' (Elon Musk in Cellan-Jones, 2014, January 2). When in fear, people become narrow-minded and tend to make Type I errors (Mujica-Parodi et al., 2017) (Fig. 2), rejecting H0 when in fact they sound a false alarm: There is no signal, no cue to a robot around.

When in fear, people's heuristic is to be better safe than sorry and look for confirmation of their (perhaps illusory) ideas. They are convinced of their right and look for conceptual confirmation of beliefs (Ross, 1977; 2018), accepting any cue that hinges on robot presence. Because their confidence is high, criterion in Fig. 2 shifts left, (e.g., $p < .1$ as measured from the right tail). This means that people in fear do not care too much about counter-evidence and are eager to accept the H1 (and reject the H0). They zero in on the danger and focus on abnormality. Strange things are happening! Which increases Type I errors.

When people experience intense positive emotions, they likely become distracted (Dreisbach & Goschke, 2004), signal discrimination becomes weaker, and they are overly optimistic about advantageous outcomes of an event (Nygren, Isen, Taylor, & Dulin, 1996). People sometimes welcome the introduction of robots for easing their loneliness or for doing household chores. When in joy, people are prone to risky behaviors (Nygren et al., 1996) and easily make Type II errors (Mujica-Parodi et al., 2017), willing to accept the H0 (no signal) when in fact there is something the matter. There may be cues that indicate the presence of something robotic. However, because the robot fulfills a need and its functions are relevant to personal concerns, the user happily ignores the robotic side of the helper and feels the robot is like a human friend (cf. Computers Are Social Actors theory, Nass & Moon, 2000).

Because in joy, people are more relaxed, accepting, and open to explore, they do not exclude new information from known categories so quickly (cf. Fredrickson & Branigan, 2005) but are willing to change their ideas and accept new category members although those may be peripheral, not prototypical. When joyful, people care less about the soundness of their beliefs but focus on the experience itself that makes them happy. Therefore, criterion in Fig. 2 shifts right, (e.g., $p < .0001$ as measured from the right tail). This means that people in a positive mood focus on sameness and togetherness rather than difference and alienation. They are not eager to accept the H1 (therefore, $p < .0001$) but desire the H0 to be true: Humans treat robots as if they were other humans (cf. Media Equation). There is nothing strange to loving a robot, which increases Type II errors.

In all, when the intensity of negative emotions is high, beliefs are willingly confirmed, and people tend to make Type I errors or call out for 'false alarms.' They do not look into the data but project their conceptualizations onto the world. What the majority thinks reality is about, fearful people say is all a fiction (Fig. 2). When positive emotions are dominant, it does not matter that the robot disconfirms beliefs because it fulfills a relevant need. People tend to make Type II errors or allow 'misses' because they pay no attention or lack the sensitivity to detect the distinction, taking fiction for real (Fig. 2). Obviously, however, what is real, not real, true or false all depends on one's conception of 'reality,' an issue we will address in the section entitled *Epistemics of the Virtual*.

4. Result: Logical fallacies and rhetorical tropes

When the relevance of an issue is high, think of personal relationships or other important social topics, people expectedly become

emotional. Emotionality, however, whether positive or negatively tinged, most likely impairs the capacity for logicality; because in our conception, the affective limbic route would shortcut cortical reflection. Blanchette and Richards (2004) found that the probability to draw invalid conclusions from emotional statements was significantly higher compared with neutral contents. Particularly the common logical fallacies of 'affirming the consequent' (*ex-consequentia*) and 'denying the antecedent' (*inverse error*) occurred more frequently in response to emotional compared with neutral statements (ibid.).

Even when taking precedence, the limbic system still is into contact with the neocortex and is reflected upon. As said, there are two main streams of problem-solving strategies of a reflective kind: logical reasoning or 'intelligence' and associative combination-making or 'creativity.' In the next subsections, we forward our ideas of what results from reasoning and creativity when dealing with false alarms that arise from negative emotions and signal misses that come from positive ones.

4.1. Ex-consequentia

The opening part of this subsection is about what most cognitive scientists with knowledge of the reasoning literature would call 'affirming the consequent' (Blanchette & Richards, 2004; Magnani, 2015) but by its Latin name. We posit that when fear causes false alarms that are addressed through logics, *ex-consequentia* reasoning may ensue (Fig. 1). This is a new hypothesis, which we want to explore theoretically before submitting it to an empirical test. In a strictly logical sense, the problem with an *ex-consequentia* fallacy is that, consistent with abductive reasoning (Magnani, 2015), it concludes the antecedent from a conditional and its consequent:

If A then B

B

Then A

For mostly pragmatic reasons, people do not rely on purely deductive inference alone as normative logicians would have it (Evans, 2002). People tend, for example, to infer external threats from internal psychological states. They believe: "If I feel anxious, there must be danger" (Engelhard & Arntz, 2005):

If there is danger, I feel anxious
I feel anxious
Therefore, there must be danger

More specifically:

If AI is dangerous, I fear it
I fear AI
Therefore, AI must be dangerous

When a robot falls of a table, is dropped, or is maltreated by its user, people tend to feel sorry for the machine (Konijn & Hoorn, 2016):

If an agency has pain, it says 'ouch'
Robot says 'ouch'
Therefore, the robot is in pain

Personal feelings and individual impressions often are used as signals to infer intrinsically 'real' psychological states, also in virtual others. Konijn, van der Molen, and van Nes (2009) found that viewers assumed genuine emotional states in soap characters because the viewers themselves were moved by the scene: "If I feel, it must be real." Translating this finding into *ex-consequentia* format:

If I see real emotions, I feel them too (empathy)
I feel emotions
Then the emotions I see must be real (misplaced empathy)

That *ex-consequentia* is a fallacy is shown by raising counter-examples (cf. Blanchette & Richards, 2004). With respect to the fear of AI, for instance, it may be that AI is not dangerous but that bad people are misusing it (i.e. people misuse other tools as well). The fear may not be induced by AI itself but by dystopic misrepresentations in Hollywood media fare (Broadbent, 2017). It may be that the fear is not induced by the AI but that AI is misinterpreted as its source.

Ex-consequentia may follow from false alarms, trying to infer the cause from the effect. For example, a person blows her nose, so she has a cold (false alarm). Someone sounds synthetic so it must be a robot (but it was Stephen Hawking, Cellan-Jones, 2014, January 2). Fear may lead to false alarms and *ex-consequentia* rhetoric.

4.2. Fp-type metaphor

Magnani (2015) makes the case that *ex-consequentia* reasoning corresponds to abductive inference, seeking causes for phenomena that otherwise remain unexplained. He deliberately connects this problem-solving style to creative processes, which make up for knowledge scarcity in situations of sparse data.

We make the case that when false alarms are used creatively, metaphors of the form 'Juliet is the sun' occur, which, for lack of better words, we name an *Fp-type* of rhetorical trope after 'false positive.' We do *not* claim that all false positive statements are necessarily metaphors. As a new hypothesis, we do posit that a metaphor or simile such as 'Juliet is the sun' or 'This man acts like a robot' have a false alarm in them and that to understand this, we should look into Frege (1892) distinction between reference and sense.

Note that it is infeasible to incorporate the huge volume of work in natural language semantics and pragmatics of metaphor that has occurred since the pivotal work of Frege. There have been substantial developments since of which we can only discuss a fraction here.

For instance, the work of Mashal and Faust as overviewed in Faust (2012, pp. 432–433) provides a signal-detection account of metaphor comprehension. Humphrey, Bryson, and Grimshaw (2010) looked into signal sensitivity to detect aptness of a metaphor in a longer sentence (cf. Lerche, Christmann, & Voss, 2018). Xie and Zhang (2014) and Huang, Xie, and Wang (2018) studied perceptional discrimination as primed by different types of metaphor. To identify metaphors in text analysis, a host of features, properties, and attributes guide the classification of metaphors; as reviewed by Shutova, Sun, Gutiérrez, Lichtenstein, and Narayanan (2017). However, this body of work may focus on the location of metaphor-comprehension processes in the brain (Faust), appropriateness of a metaphor in a context (Humphrey et al.), visual detection (Xie & Zhang, 2014), similarity estimates in text mining (Shutova et al.), or designate the notion of 'false alarm' itself as a metaphor (Juhasz & Sarbin, 1966); it does not employ detection flaws to typify certain kinds of metaphor, as we do here.

As said, we went back to the basic semantic logics of language that underlies all of the metaphor literature in which a 'target domain' is compared to a conventionally unrelated 'source domain.' In metaphor theory, many names address the same thing: Dependent on the theory, 'topic' may be called '*primum comparandum*,' 'principal subject,' or 'tenor' and the imagery is sometimes called '*secundum comparandum*,' 'secondary subject,' 'marginal meaning,' or 'vehicle,' etc. We will comply with Frege (1892) distinction between 'reference' and 'sense.'

The reference of a word is to something existing in the real world (according to the observer's beliefs). The 'sense' of the word is its presentation form. The famous example is that the second planet from the sun may be called 'Venus' as well as 'the morning star' and 'the evening star' (although it is not a star at all). All three expressions are synonyms for that one particular planet. The 'sense' of a word emphasizes a particular aspect of the entity it refers to, conjuring up an image, a 'mock thought' or 'Scheingedanke' about that entity.

With 'Juliet is the sun,' something similar happens. Presupposing that the proper name refers to a girl (that once lived), 'sun' is the image,

the 'mock thought' that highlights an aspect of the girl that the speaker admires (e.g., Juliet is bright). Juliet (supposedly) has reference to an entity, whereas with regard to the features of that entity, 'sun' has sense alone. In the case of an *Fp-type* of metaphor or simile (This man acts like a robot), the topic or focus of the comparison coincides with the reference (a specific man), while the imaginative part coincides with the sense (inflexible person): The noise is supplemented by a fake signal (Fig. 2).

4.3. Inverse error

Apart from false alarms that come out of fear, misses may occur as well, perhaps because a person was overwhelmed by joy. A logical fallacy that likely may arise from missed signals is the *fallacy of the inverse* (Fig. 1). The *fallacy of the inverse* or *inverse error* denies the antecedent (Blanchette & Richards, 2004) while inferring the inverse from its original statement:

If A then B
Not A
Then not B

For example:

If I see machine-like features, the agency is a robot
I do not see any machine-like features (missing the signal)
Then the agency is not a robot

To its conclusion, one can raise counter examples to show that an inference is a fallacy: The robot has autonomous systems that simulate human behavior very well. There may be an operator in the background that handles the machine. Forwarding counter-examples shows that there is a need for an ontology or knowledge base to draw the examples from. We will return to this issue later in the section *Reference - sense*.

4.4. Fn-type metaphor

When missed signals are used creatively, metaphors transpire such as 'Robots are human too.' These we will call Fn-type of tropes after 'false negative.' Different from *Fp-type*, *Fn-type* of expressions focus on the word that carries sense alone as the topic (robot) and compare that to the word that has reference to the observer's conception of reality (human beings): The missed signal is filled up with noise (Fig. 2).

5. Fundamental attribution errors

In the previous, we discussed four expressions that are non-literal and false, two coming from detection faults combined with flawed reasoning (*ex-consequentia* and *inverse error*) and two from creativity (*Fp-* and *Fn-type* metaphors). In this section, we argue that these four allow for fundamental attribution errors that people make with respect to their fellow humans as well as to robots, avatars, and autonomous AI.

In interaction with computers such as a voice agent in a navigation system, people consistently show self-serving biases, attributing positive results to their own doing ("I found it!") and when things go wrong, accusing the computer of poor performance ("You did not warn me in time!") (Moon, 2003; also Groom, Chen, Johnson, Kara, & Nass, 2010). According to Caporael (2006), the robot as anthropomorphism (Section *Anthropomorphism, Robotomorphism*) may be understood from the so-called 'fundamental attribution error.' Even the simplest of robots so Caporael argues supposedly 'wants' to navigate a room or wishes to chat, while users assume it has an internal disposition that 'motivates' the robot's behaviors.

Suppose a user is treated courteously by her robot. The user is delighted by the well-mannered robot and tends to attribute the robot's behavior to its 'personality,' as people do to other people. S/he may believe the robot really likes him/her and has a friendly character. That user does not think that the robot follows mere protocol, implemented by a skilled programmer, and that it knows politeness nor friendliness. When the robot does not look at her during conversation, that user will not think the robot runs out of power, has a broken camera eye, or that the video overheated its CPU. It is just that the robot 'does not feel like it today' (cf. Moon, 2003). People tend to believe that what the robot does reveals what goes on inside while ignoring external factors that explain its behaviors (such as a person that has the machine in remote control). This is what social psychologists call a 'fundamental attribution error' (Caporael, 2006; Ross, 1977), inferring internal qualities from external cues.

Fundamental attribution errors are related to signal detection in that misses and false alarms both facilitate their occurrence albeit of a different kind. Imagine a user in a Turing Test with an avatar on screen, half the trials of which are handled by a human being and the other half by an AI. On certain trials, participants are sitting across a human being although they think they interact with the AI (false alarm, it is a confederate). However, all scripts and schemas of computers and machines become active. In that situation, the participant attributes internal robot qualities (rigid, cold) to the human confederate as based on the external performance of the avatar. In other words, false alarms to robot cues lead to fundamental attribution errors derived from the (non-social) schemas of machines. That the user makes such mistakes is likely denied. They may seek the cause in the drop-down menus of the interface, which avoided human emotional behavior of the confederate to come to full expression.

On other trials during that Turing Test, participants sit across an AI but believe they interact with a human being (miss, it is an AI). All scripts and schemas for human interaction remain active (cf. Nass & Moon, 2000). Thus, the participant attributes human qualities to that robot (kind, warm) as based on the external performance of the avatar. Missing the cues to a robot leads to fundamental attribution errors drawn from *human* social schemas and templates. Users will likely believe they are excused for their mistakes, for example, because no one told them when they would talk to a machine or not (which is the very idea of a Turing Test).

6. Anthropomorphism, robotomorphism

Humans are inclined to attribute human emotions to animals in order to understand their behavior (Darwin, 1875, 2002). They do so for robots as well. Based on the level of human-like appearance (e.g., eyes), for example, people tend to overestimate what robots can do functionally (i.e. it can look around and is aware of its surroundings) (Haring, Watanabe, Velonaki, Tossell, & Finomore, 2018). This tendency to project human characteristics to living as well as non-living, virtual as well as real entities is called anthropomorphism (Anzalone, Boucenna, Ivaldi, & Chetouani, 2015).

According to Epley, Waytz, Akalis, and Cacioppo (2008) and Epley et al. (2007), anthropomorphism is a means to make inferences about an unknown entity by applying knowledge from known agencies. In our terms, the missed signal is filled up with noise. To do so, people should have the motivation (e.g., have fun), have a need (cf. loneliness), or the unknown entity should show certain similarities (cf. humanoids).[5]

People attribute internal human qualities to other humans as based on external behaviors of the other, which we call a fundamental

[5] Robot developers tend to design robots after humans ('humanoids') from the presumption that people like them better and treat them better that way (cf. Złotowski, Proudfoot, Yogeeswaran, & Bartneck, 2015). For instance, human-like robots were punished less and praised more than robots that were less human-like (Bartneck, Reichenbach, & Carpenter, 2006). Human-likeness does not just pertain to outer appearance but also to the robot's autonomy, communication, (emotional) behavior, intelligence, and predictability (Złotowski et al., 2015).

attribution error. When people attribute internal human qualities to objects, concepts, or non-human agencies (real or virtual) as based on the external behaviors of those entities, then we have a fundamental attribution error coming from poor signal detection, particularly missing the cue to non-humanness, leaving the schemas of human social behavior intact. This we call an anthropomorphism, which is a specific fundamental attribution error as it applies to non-human entities and is based on missing (perhaps deliberately ignored) signals to non-humanness.

When based on false alarms, schemas of machine behaviors become active (possibly internalized from film and other media fare) and 'the other humans' are alienated by assuming internal non-human, non-social, mechanisms and automated behaviors in them (cf. bureaucrats and technocrats). This we may call 'robotomorphism.'

Taken in unison, false alarms regarding cues to robots produce *ex-consequentia* fallacies and *Fp-type* of metaphors that activate behavioral schemas and scripts reminiscent of machinery that then are attributed falsely to humans: robotomorphism (Fig. 1). Missed cues to robots produce *inverse errors* and *Fn-type* of metaphors so that current human behavioral scripts are not deactivated, which then are attributed falsely to robots: anthropomorphism (Fig. 1).

7. How to tell a robot: A four-pronged approach

A user may instruct his robot while using fallacious syllogisms and rhetorical tropes, all in one sentence: 'If you don't look at me, you metal monkey, you can't pay attention to what I say.' The fallacy is denying the antecedent (*inverse error*) because the robot's microphones work perfectly without looking at the speaker. Metaphorically, the user compares the robot to a monkey for its foolishness. What should the robot do to bring back such utterances to statements it can execute?

People invented programming languages (i.e. Fortran) and graphical interfaces ('click') to convey unambiguous instructions to the machine but during natural-language interactions between humans and robots, such interaction modalities may not likely be preferred (cf. Angleraud et al., 2019). How should a robot deal with emotionally aroused users that make logical mistakes or use rhetorical tropes such as metaphors and similes? Should the robot tell its user wrong all the time and in doing so, add to the anger or spoil the fun?

One moral demand is that a social robot ought to tell the truth (deontics). That means that the robot should point out mistakes to its user. Then again, a communication demand is that the robot is polite and tactful. Sometimes, these demands are at odds. Internally, a robot needs logical language to execute its tasks but attribution errors are often based on logically invalid reasoning and detection wrongs. How to avoid that the user becomes lectured by its robot as Doctor McCoy was by Mr. Spock: "I find your arguments strewn with gaping defects in logic" (*Star Trek II: Wrath of Khan*).

We propose four measures for a robot to deal with logical fallacies and figures of speech. For its ontology, it needs to discern reference from sense (Frege, 1892) (Section *Reference - sense*). To prove an argument is false, it may apply tableau reasoning (Beth, 1955) (Section *Tableau reasoning*). To keep the robot from responding 'error' and to suspend disbelieve, it should assess its user's non-literal and false utterances under an intensional operator that is authorized by Grice (1975) Maxim of Quality (see the section *Maxim of Quality*). To deal with an imaginative world, mock thought, or 'Scheingedanke' conjured up by the user's fallacies and tropes, the robot should run two times an 'Epistemics of the Virtual' (Hoorn, 2012), once for its user and once for itself (see the section *Epistemics of the Virtual*).

7.1. Reference – Sense

Philosophically, an ontology is the study of what is and is not in the world, what entities exist (God, Martians, dark matter), what not, and what categories they are a member of. Each ontology is observer-dependent. Translated into a computer system, an ontology is a formal knowledge base, representing entities (objects, concepts) and their relationships that describe some part of reality (e.g., the user's family). With an ontology installed, the robot can make inferences about the composition of that domain, for example, if A is the child of a man's sister, he must be A's uncle.

For a robot to grasp the semantics of fallacies and rhetorical tropes, the ontology it works with should be structured in categories (e.g., animals), exemplars (e.g., snakes), and features (e.g., head, long, tail, no legs). As soon as a category mismatch happens (Juliet is an exemplar in human beings but star is not), the robot knows that one of the terms has reference (within category) and that the other probably has sense alone (out of category). This way, the robot knows it is dealing with a non-literal utterance.

However, ontologies are observer-dependent. They sometimes need to be matched, reconciled, or 'harmonized' (e.g., Hildebrandt, Törsleff, Caesar, & Fay, 2018). The robot's ontology (O_R) is just one of two ontologies that are matched against each other. The other is the ontology of the human (O_H), which may differ from that of the robot. For instance, the robot may hold numbers for physical features of objects, whereas the human classifies them under ideas. Or the human classifies 'robots' under the category of human beings whereas the robot does not. The robot may believe that ghosts do not exist 'but my user does.' Like this, the robot can work with more perspectives or worldviews. Thus, each item in the ontology has a probability attached that indicates the strength of belief that the agency thinks something is true. Thus, the robot holds one ontology O_R for itself, containing the things the robot believes in and it holds one ontology O_H for its user, containing the things the robot believes that the user believes in. Like this, the robot has 'theory of mind.'

7.2. Tableau reasoning

Next, the robot determines whether it deals with a fallacy and for the robot to analyze whether the user went wrong, it may use tableau reasoning. Tableau reasoning is a predicate logic to (dis)prove a conclusion based on the ontology or knowledge base of an individual (Beth, 1955; 1956). The aim of tableau analysis is to prove that statements imply other statements (implications) in view of the rules of meaning of the statements' connectives or quantifiers in first-order logic.

Through tableau reasoning, the analyst proves a statement by refutation of the contrary (i.e. for A to be true, not-A must be false). For every single statement, then, the possibilities of alternative truths are tested, trying to prove the falsehood of the implications (i.e. the negation of the ensuing statements). The semantics of the ontology may be, for example, that the category of human beings has Juliet as an exemplar but not robots. There also is a certain belief attached to each entity in the ontology. Certain entities are more probable (e.g., human beings have hands) than other (e.g., human beings have fish scales). The semantics of that ontology then transforms the list of alternative truths into a tree of simpler statements so to find a contradiction for each branch. To arrive at the simpler statements, reasoning rules apply (Fig. 3). If contradictions are found between branches, it may be concluded that the original statements and the negation of their implications cannot be true all at once so that it follows that the original statement indeed implies its conclusions.

Underlying a user's statement is a certain belief in O_H of which the statement is a logical consequence. The robot tests if the statement as logical consequence of the user's beliefs is true given the information in the ontology O_H. In making use of the rules in Fig. 3, the robot then divides the logical formula that contains that knowledge into different components with their logical consequences. Each logical sequence is transformed into an NNF (negation normal form) in which a negation sign ($\neg$) is placed in front of the statement under scrutiny. This is done to check if all the components from O_H are in contradiction with all the

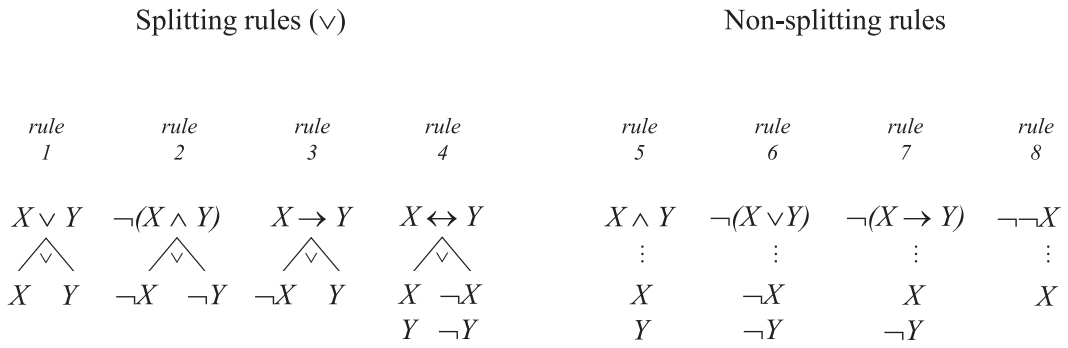


Fig. 3. Rules of tableau analytics.

components of the logical sequence. When all sub-components of O_H and the logical consequence are in contradiction, the robot regards the logical consequence as proven true.

Next, we illustrate the procedure with an example. It is an *ex-consequentia* reasoning structure, coming from a false alarm in detection (Fig. 1). Suppose the user says: "I feel anxious, so there must be danger" (Engelhard & Arntz, 2005). Translated into logics, the robot holds that:

anxiety → danger

From the O_H that the robot also keeps, the robot knows that the user knows that the reverse is true as well: Being in danger causes anxiety.

danger → anxiety

Based on O_H, the epistemic logics of the robot is that this human knows that danger causes anxiety but that the user inversed that relationship. Let:

B = Believes
K = Knows
O = Ontology ('knowledge base')
R = Robot
H = Human

and let O_R contain:

$$\mathbb{B}(R)\trianglerighteq\mathbb{B}(H(danger))\rightarrow\mathbb{K}(H(anxiety)) \quad (1)$$

Equation (1) is the worldview that the robot holds as 'standard,' the one most agencies in its community share. From the user's statement, however, the robot observes that this human also seems to believe that anxiety is causing danger:

$$\mathbb{B}(R)\trianglerighteq\mathbb{K}(H(anxiety))\rightarrow\mathbb{B}(H(danger)) \quad (2)$$

Statement (2) seems a deviant thought compared to (1) and may now be tested through tableau analytics. The robot works from O_R and starts out from assuming its truth. The logical consequence of (2) contains the ontology from the current O_H; it is a new 'hypothesis' of which the robot is unsure about its truth. With the rules of tableau reasoning installed (Fig. 3) (Beth, 1955; ,Beth (1956)), the robot may try to solve the *ex-consequentia* fallacy the user apparently makes in (2).

Logical consequences
apply
$(O_R)\mathbb{B}(R)\trianglerighteq\mathbb{B}(H(danger))\rightarrow\mathbb{K}(H(anxiety))$ $X\rightarrow Yr3$
a = anxiety
d = danger
$\neg\mathbb{B}(R)\vee(\mathbb{B}(H(d))\rightarrow\mathbb{K}(H(a)))$ $\neg X\vee Y$
$\neg\mathbb{B}(R)\vee(\neg\mathbb{B}(H(d))\vee\mathbb{K}(H(a)))$ $\neg X\vee(\neg y_1\vee y_2)$
$(O_H)\mathbb{B}(R)\trianglerighteq\mathbb{K}(H(anxiety))\rightarrow\mathbb{B}(H(danger))$ $X\rightarrow Yr3$
$\neg\mathbb{B}(R)\vee(\mathbb{K}(H(a))\rightarrow\mathbb{B}(H(d)))$ $\neg X\vee Y$
$\neg\mathbb{B}(R)\vee(\neg\mathbb{K}(H(a))\vee\mathbb{B}(H(d)))$ $\neg X\vee(\neg y_2\vee y_1)$
$\neg(\neg\mathbb{B}(R)\vee(\neg\mathbb{K}(H(a))\vee\mathbb{B}(H(d)))$NNF $\neg(\neg X\vee(\neg y_2\vee y_1))$
$\mathbb{B}(R)\vee\neg(\neg\mathbb{K}(H(a))\vee\mathbb{B}(H(d)))$ $X\vee\neg(\neg y_2\vee y_1)$
$\mathbb{B}(R)\vee(\mathbb{K}(H(a))\vee\neg\mathbb{B}(H(d)))$ $X\vee(y_2\vee\neg y_1)$
Proving or disproving the consequence
(O_R)
1 $\neg\mathbb{B}(R)\vee(\neg\mathbb{B}(H(d))\vee\mathbb{K}(H(a)))$ $\neg X\vee(\neg y_1\vee y_2)r1$
$2|\alpha$ from $1\neg\mathbb{B}(R)$ $\neg X$
3 $\neg\mathbb{B}(H(d))\vee\mathbb{K}(H(a))$ $\neg y_1\vee y_2r1$
$4|\alpha$ from $3\neg\mathbb{B}(H(d))$ $\neg y_1$
5 $|\alpha$ from 3 $\mathbb{K}(H(a))$ y_2
(O_H)
6 $\mathbb{B}(R)\vee(\mathbb{K}(H(a))\vee\neg\mathbb{B}(H(d)))$ $X\vee(y_2\vee\neg y_1)r1$
$7|\beta$ from 6 $\mathbb{B}(R)$ Contradiction with $2|\alpha$ X
8 $(\mathbb{K}(H(a))\vee\neg\mathbb{B}(H(d)))$ $(y_2\vee\neg y_1)r1$
$9|\beta$ from 8 $\mathbb{K}(H(a))$ No contradiction with $5|\alpha$ $y2$
$10|\beta$ from $8\neg\mathbb{B}(H(d))$ No contradiction with $4|\alpha$ $\neg y1$

Because through tableau analysis (Fig. 4), the robot does not find a contradiction in all cases, it cannot conclude that (2) $\mathbb{B}(R)\trianglerighteq\mathbb{K}(H(anxiety))\rightarrow\mathbb{B}(H(danger))$ is true based on (1) $\mathbb{B}(R)\trianglerighteq\mathbb{B}(H(danger))\rightarrow\mathbb{K}(H(anxiety))$. The robot has not found proof to accept the human statement and knows that the human drew an incorrect conclusion (i.e. *ex-consequentia*). In conventional systems, an error message would be outputted (ID10T).

Let us now turn to handling the *inverse error* structure:

If A then B
Not A
Then not B.

If 'cues to machine' then agency is a robot
No cues to machine detected (miss)
Then agency not a robot

Let in the same ontology O_R as before:

A = Agency
cm = cues to machine
Then,

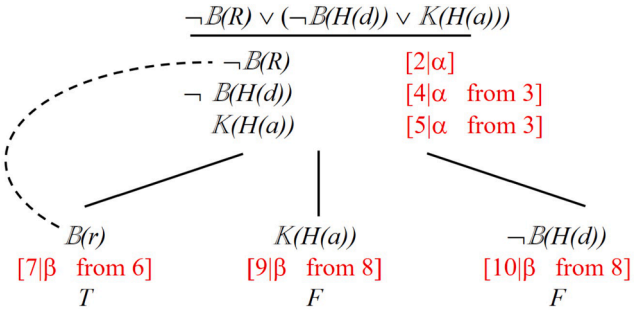


Fig. 4. Tableau analysis of an *ex-consequentia* fallacy (T = true, F = false).

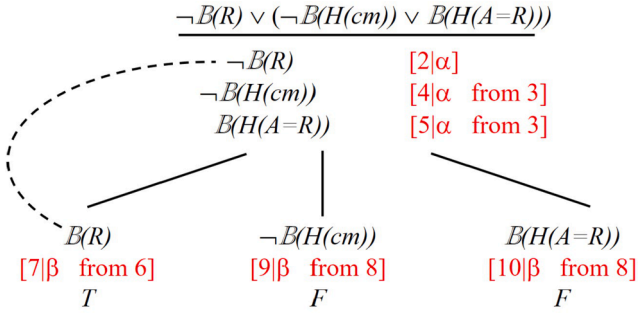


Fig. 5. Tableau analysis of an *inverse error* (T = true, F = false).

Logical consequences
apply
$(O_R)\mathbb{B}(R)\trianglerighteq\mathbb{B}(H(cues_{to_{machine}}))\rightarrow\mathbb{B}(H(Agency = Robot))\quad X\rightarrow Y\quad r3$
$\neg\mathbb{B}(R)\vee(\mathbb{B}(H(cm))\rightarrow\mathbb{B}(H(A=R)))\quad \neg X\vee Y\quad r3$
$\neg\mathbb{B}(R)\vee(\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R)))\quad \neg X\vee(\neg y_1\vee y_2)$

Given its ontology O_R, the robot now tests whether it believes it is true that 'when a human does not believe that there are cues to a machine, an agency is not a robot' (Fig. 5).

$(O_H)\mathbb{B}(R)\trianglerighteq\neg\mathbb{B}(H(cm))\rightarrow\neg\mathbb{B}(H(A=R))\quad X\rightarrow Y\quad r3$
$\neg\mathbb{B}(R)\vee(\neg\mathbb{B}(H(cm))\rightarrow\neg\mathbb{B}(H(A=R)))\quad \neg X\vee(\neg y_1\rightarrow\neg y_2)\quad r3$
$\neg\mathbb{B}(R)\vee(\mathbb{B}(H(cm))\rightarrow\neg\mathbb{B}(H(A=R)))\quad \neg X\vee(y_1\vee\neg y_2)$
$\neg(\neg\mathbb{B}(R)\vee(\mathbb{B}(H(cm))\vee\neg\mathbb{B}(H(A=R))))$NNF $\neg(\neg X\vee(y_1\vee\neg y_2))$
$\mathbb{B}(R)\vee\neg(\mathbb{B}(H(cm))\vee\neg\mathbb{B}(H(A=R)))\quad X\vee\neg(y_1\vee\neg y_2)$
$\mathbb{B}(R)\vee\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R))\quad X\vee\neg y_1\vee y_2$

Proving or disproving the consequence
1 $\neg\mathbb{B}(R)\vee(\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R)))\quad \neg X\vee(\neg y_1\vee y_2)\quad r1$
2|α from 1 $\neg\mathbb{B}(R)\quad \neg X$
3 $\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R))\quad \neg y_1\vee y_2\quad r1$
4|α from 3 $\neg\mathbb{B}(H(cm))\quad \neg y_1$
5|α from 3 $(H(A=R))\quad y_2$
(O_H)
6 $\mathbb{B}(R)\vee(\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R)))\quad X\vee(\neg y_1\vee y_2)\quad r1$
7|β from 6 $\mathbb{B}(R)$ Contradictionwith2|α X
8 $\neg\mathbb{B}(H(cm))\vee\mathbb{B}(H(A=R))\quad \neg y_1\vee y_2\quad r1$
9|β from 8 $\neg\mathbb{B}(H(cm))$ No contradiction with 5|α $\neg y1$
10|β from 8 $\mathbb{B}(H(A=R))$ No contradiction with 4|α $y2$

To prepare the robot for conversational repair, we believe that each *ex-consequentia* fallacy and *inverse error* can be analyzed in the above manner. This, of course, is a hypothesis that should be tested through theorem proving in an automated corpus analysis while the results are checked by expert logicians.

7.3. Maxim of Quality

As demonstrated in the above, humans use logical fallacies in their communication. Grice (1975) formulated a number of conversational maxims of which the Maxim of Quality is most useful to our purposes. Maxim of Quality approaches the speaker from the assumption that the user intends to speak the truth and provides honest evidence to support his or her statements (with a subtle difference, also Lepore & Stone, 2015, p. 2). For example, Jwalapuram (2017) used Gricean maxims to evaluate dialogs between users and chatbots. The Maxim of Quality was of particular interest to check if the conversation dialog system was faithful to the factual knowledge provided to it. In a formalization to extract rules from natural language, Sorower et al. (2011) inverted Gricean maxims and implemented these in Makovian logics. This way, the computer could learn and employ conversational rules but the system had no way to 'excuse' the user for making a mistake.

In the case of a fallacy, however, the speaker merely makes an error in the form, not in contents. Strictly speaking, with the Maxim of Quality the robot counters one fallacy with another. Maxim of Quality (MQ) is a fallacy *ad ignorantiam* but pragmatically needed for the robot to 'maintain the H0′ so that the user is 'innocent until proven guilty.' In applying an MQ operator to a logically fallacious proposition, human-robot conversations are rescued from the Spock syndrome of eternal bickering.

With MQ activated, the robot then should check three things with its user. Robot should ask about the danger: "So it's unsafe you think?" It should check the user's anxiety: "You're really anxious, aren't you?" And it should verify the logical consequence: "Do you believe that your fear indicates a threat?" The robot could also just repeat the whole statement "You feel anxious so there must be danger?" and then include all three automatically.

If all answers are 'yes,' the robot now believes *B(R)* that the user knows s/he is anxious *K(H(a))* and that the user believes there is danger *B(H(d))*.

However, we suggest that after proving fallacy, the robot should not output an error message. Instead, the robot's uncertainty whether the human really means that anxiety causes danger may be expressed under MQ as a special instantiation of the 'possibly true' operator ◊ in intensional logics:

$$\mathbb{B}(R)\trianglerighteq\Diamond\mathrm{MQ}(\mathbb{K}(H(anxiety))\rightarrow\mathbb{B}(H(danger))) \tag{3}$$

The robot then may update its knowledge about the user (O_H) such that:

If user senses anxiety (B)
And if B then A is **necessarily** (□) false to robot (false = *F*)
And Maxim of Quality (MQ) is **possibly** true
Then A (danger) is **possibly** true to user

$$\mathbb{B}(R)\trianglerighteq\Box(\mathbb{B}(H(danger))\rightarrow\mathbb{K}(H(anxiety)))\wedge$$

$$\mathbb{B}(R)\trianglerighteq\Diamond\mathrm{MQ}((\mathbb{K}(H(anxiety))\rightarrow\mathbb{B}(H(danger)))F)\rightarrow\Diamond\mathrm{MQ}\ (\mathbb{B}(H(danger))) \tag{4}$$

The fallacy fail-safe formula with epistemics and modalities in (4) says that robot believes it is necessary true that if user believes there is danger, then user senses (knows) s/he is anxious and robot also believes that possibly user is truthful if user knows s/he is anxious and wrongfully (*F*) infers danger from it, so that possibly the user believes there is danger.

The fallacy fail-safe formula also works for *inverse error:*

$$\mathbb{B}(R)\trianglerighteq\Diamond\mathrm{MQ}(\neg\mathbb{B}(H(cm))\rightarrow\neg\mathbb{B}(H(A=R))) \tag{5}$$

If user believes that there are no cues to a machine in some agency (¬A)
And if ¬A then ¬B is **necessarily** false to the robot
And Maxim of Quality is **possibly** true
Then ¬B (agency is no robot) is **possibly** true to user

$$\mathbb{B}(R)\trianglerighteq\Box(\mathbb{K}(H(cm))\rightarrow\mathbb{K}(H(A=R)))\wedge$$

$$\mathbb{B}(R)\trianglerighteq\Diamond\mathrm{MQ}((\neg\mathbb{B}(H(cm))\rightarrow\neg\mathbb{B}(H(A=R)))F)\rightarrow\Diamond\mathrm{MQ}(\neg\mathbb{B}(H(A=R))) \tag{6}$$

With (6), the robot believes it is necessary true that if user knows there are cues to a machine in some agency, then the user believes the agency is a robot. Robot also believes that possibly user is truthful if user states s/he does not see cues to a machine and wrongfully infers the agency cannot be a robot. Thus, if the user believes there are no cues to a machine, then it is possible that the user believes that the agency is not a robot (and treats it as a human being). Again, the fallacy fail-safe formulas are hypothetical in their current form and should be verified in user tests to see if the robot indeed is capable of conversational repair this way.

7.4. Epistemics of the Virtual

Gricean maxims for conversational repair do not take background knowledge into consideration (Bernsen, Dybkjaer, & Dybkjaer, 2014, p.

121-122). Yet, with the introduction of the $\Diamond_{MQ}$ operator, the robot enters the realm of possibilities rather than applying rules to a known ontology with 'historic' information. Put differently, by inserting $\Diamond_{MQ}$, we actually demand from the robot to open up to a fictional or mistaken account of the world (cf. mock thoughts or 'Scheingedanke') (Fig. 6, right). To understand the 'sense' of a message, the robot must assume that under certain conditions in a certain context with certain parameter settings, there are possible worlds in which an utterance may be true.

In our lab, we develop a software called *EpiVir*,[6] which is short for Epistemics of the Virtual (Fig. 6, Hoorn, 2012). *EpiVir* is a system that builds up and changes an ontology, according to incoming information. *EpiVir* acknowledges that it deals with a *representation* of the physical world, which it calls 'Reality' (i.e. its own particular representation of the world). Within that specific Reality, *EpiVir* attributes a truth value to incoming information in the range [0–1] with possibilities for 'partial truths' [0.5, 0.3, etc.]. Truth is attributed to statements or observations, according to beliefs that are relatively stable (e.g., God exists, Martians do not. Superposition exists, quantum foam doesn't). Truth values are not fixed but can change under the influence of new data. Information does not leave the system but lies at a higher or lower level of activation, depending on the goals and concerns of the agency. This is how *EpiVir* builds up its database with probabilities on the ontological status of its contents (true, possible, false).

One segment of the database is categorized as Fiction (Fig. 6, right). They are the entries in the database that range from [0.5–0], from possible to false. They may refer to motion pictures, theater acts, as well as exaggerations, mistakes, and plain lies. In Frege (1892) terms, it is the realm of sense rather than reference.

There is a continuous loop, checking new input against known categories. This is a rough, low-level template check for continuity, whether stored concepts are still in line with sensor data, whether the ontology is still up to date. If information enters that differs too much from the template in terms of signal detection, a validation process becomes more highly activated. This is an effort-intense, precise, and detailed epistemic appraisal of the deviant stimulus. Ontological classification is concept-driven, whereas epistemic appraisal is more data-driven. Both processes run in parallel but one may be at a higher activation level than the other.

From this, a number of assessments is made with respect to the ontological status of the information in the database. Information may be more or less true, is part of fiction or reality, and seems more or less realistic. This way, the system can deal with someone staring out of the window [Reality], saying: "It may be true but I can't believe my eyes" [unrealistic]. Or someone watching a soap series on TV [Fiction], saying: "Yes, I know this is not true, but such is life" [realistic].

When the robot encounters a logical fallacy, it initially will place a syllogism such as (2) in the area of 'false,' an unrealistic fictitious assessment of reality at the most. By using a fallacy fail-safe formula such as (4), the robot now may move the 'contained' fallacy into the area of truth and reality because the robot's rendition is correct that its user believes the user's rendition is correct although in the robot world, the user's rendition is not.

When the robot encounters a metaphor or simile, it should accept a proposition that is not literally true (cf. Grice, 1989, p. 34), which means that the robot should be capable of dealing with possible worlds. In Epistemics of the Virtual (Hoorn, 2012), the fiction module is opened through 'suspension of disbelief' (cf. $\Diamond_{MQ}$), where uncertain propositions are tested for truth through empirical scrutiny (called 'epistemic appraisal'). Such propositions happen when we say 'suppose that,' 'assume that,' or 'imagine a possible world in which…'

A creative proposition such as a metaphor (e.g., a human is a machine) is counterintuitive but not implausible: Owing to topological invariance, a donut can be transformed into a coffee mug (i.e. homeomorphism) (e.g., Hung, 2016). Similarly, an orange looks like the sun because both are spheres. A chair is a table with part of the tabletop put upright. Gravity can be folded together with acceleration. Because of its disc-like shape, a coin may function as a button on a sleeve. In the same vein do people have all kinds of machine-like qualities such as the autonomous nervous system that automatically runs its programs. There are psychological scripts and social rituals that people follow as if they were programmed to do so. Humanoid robots look and behave like humans because indeed they are supposed to imitate humans.

A creative proposition comprises a conditional and intensional category attribution error (i.e. a human is not a machine, a girl is not a mermaid) from which an (over)generalization follows (i.e. all humans are machines). In Fiction $\dot{F}$, however, such an utterance is considered not necessarily false (Fig. 7).

In being a creative proposition, a simile such as 'a human is like a robot' or 'a robot is like a human' makes the following statement about the world:

> If it is possible that organisms are not machines and that if there is an organism (e.g., a human) the form of which partly resembles the form of a machine (e.g., a robot) then imagine as if it is not necessarily false that a human belongs to the machines from which follows that organisms can be like machines. (7)

To formalize (7), let X refer to the category of *Organisms* and Y to the category of *Machines* and let $\dot{\imath}$ (t and i combined) designate 'topological invariance' then (7) can be rewritten like (8) and then generalized to (9). This procedure may be used to keep a robot from responding 'error' to a user's creative utterances:

$$\text{If}\Diamond\big((\text{Organisms} \neq \text{Machines}) \wedge \{\exists x \in X | \dot{\imath}\ Human \approx \dot{\imath}\ Robot\}\big) \rightarrow$$
$$asif\big(\neg\Box\text{MQ}\neg((Human \in Machines) \rightarrow (Organisms \approx Machines))=1\big) \quad (8)$$

$$\text{If}\Diamond\big((\text{X} \neq \text{Y}) \wedge \{\exists x \in X | \dot{\imath}\ x_i \approx \dot{\imath}\ y_i\}\big) \rightarrow \dot{F}\big(\neg\Box\text{MQ}\neg((x_i \in Y) \rightarrow (X \approx Y))=1\big) \quad (9)$$

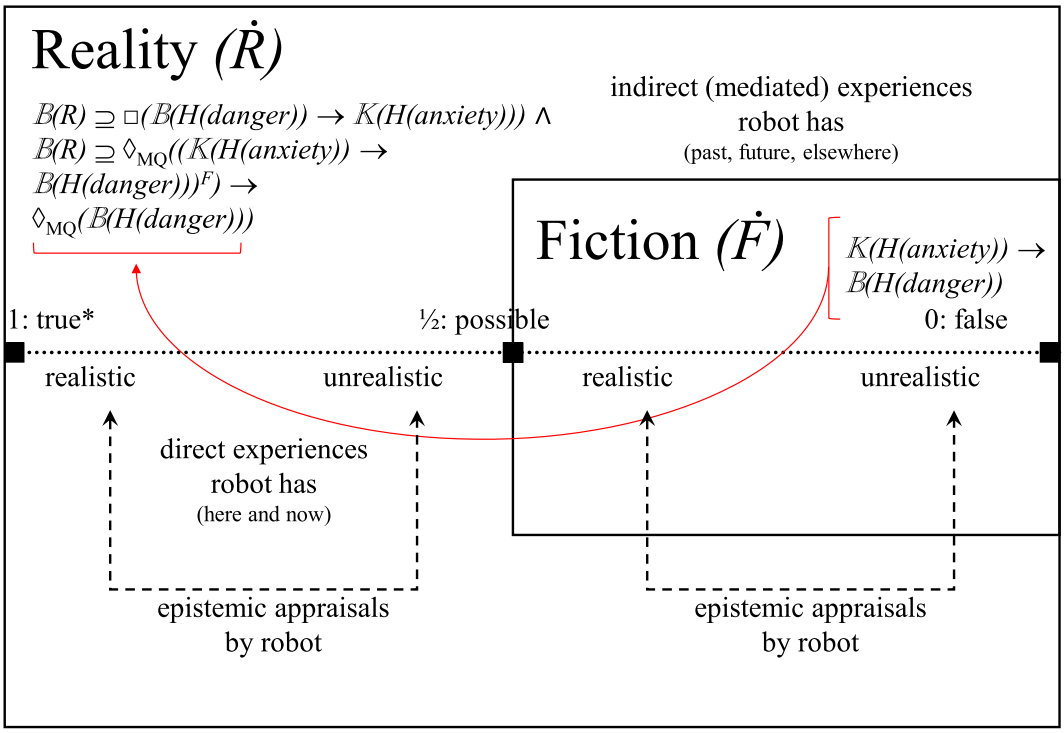
In (9), there are three conditions that underlie the acceptance of a creative proposition:

i) $X \neq Y$, as tested by concepts of reality, beliefs, or world knowledge in *EpiVir*
ii) $\{\exists x \in X | \dot{\imath}x_i \approx \dot{\imath}y_i\}$, which is tested by: 'Through topological invariance, I can demonstrate that a human is like a robot (or a robot is like a human): For instance, both have arms, legs, etc., and execute scripted behaviors.'
iii) The category-attribution error $x_i \in Y$ and the consequential (over) generalization $X \approx Y$ should be saved from rejection with the $\dot{F}$ predicate that indicates *as if*. The condition before the $\dot{F}$ predicate refers to statements about Reality $\dot{R}$ with a certain probability of being true (intensional operator $\Diamond$). After the $\dot{F}$ predicate, an imaginative world is proposed ('suppose that…') in which statements have a probability of not ($\neg$) necessarily (intensional operator $\Box_{MQ}$) being false under the Maxim of Quality (also see truth continuum Fig. 7).

This is the way we suppose the robot may deal with metaphors, similes, and other creative propositions. Obviously, this is another hypothesis that should be checked in large corpora to find examples that do not fit the proposed structure, which then should be confronted with real-life users to validate its use for conversational repair.

The $\dot{F}$ predicate may be activated by the robot when running into a logical fallacy, category mismatch, recognition of the genre or genre

[6] https://github.com/robopop/epistemics; https://github.com/robopop/docker/tree/master/epistemics



*Attribution of truth according to robot's belief system or world view (e.g., scientific, religious, or cultural)

Fig. 6. Epistemics of the Virtual.

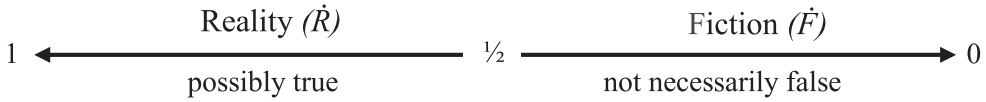


Fig. 7. Truth continuum with probabilities of truth as related to reality and fiction judgments.

attribution, for instance, when going to the movies, being in a VR environment, or hearing someone's dreams. The topological similarity of forms provides the opportunity to transgress conventional boundaries between reality and fiction and make a justifiable category-attribution error. The robot does not tell its user wrong. Instead, the category-attribution error is the impetus to the special metaphor-identification process (Hoorn, 2012, p. 106), which leads the robot to finding similarity in meaning, for example, that both human and robot are associated with agency, intelligence, and autonomy.

The overgeneralization sets up an ontology that contains fictional elements, forming the background against which subsequent utterances can be evaluated. If socially accepted, the overgeneralization may install a field of conventional metaphors (cf. Lakoff & Johnson, 2003). For instance, 'humans are like robots' may become generalized to *Organisms are Machines*.

7.5. Fail-safe protocol for fallacies and tropes

These are the steps the robot should take when encountering a logical fallacy or a category-attribution error (which occur in metaphors):

I. Run Epistemics of the Virtual *(EpiVir)* on (O_R) and (O_H)
II. Observe mismatches between a user statement and the ontology of the user (O_H) and/or the ontology of the robot (O_R)
III. Apply Maxim of Quality and suspend disbelief (start timer, set duration)
IV. Do tableau reasoning. Can tableau be closed?
- If yes, no fallacy. Go to VI
- If no, contain fallacy through (4) or (6). Go to VI

V. Do creative-proposition analysis with (8) and (9)
- If statement is a category-attribution error, classify as literally false/unrealistic in $O_R(\dot{F})$, figuratively true/realistic in $O_R(\dot{R})$. Go to VI
- If not, return error but do not output 'nonsense.' Go to VII

VI. Respond you understand, do not tell wrong, close disbelief-suspension timer
VII. Respond you do not understand, do not tell wrong, ask clarification, do not close disbelief-suspension timer

For a robot to produce an analytic tableau or creative-proposition analysis, it should maintain a knowledge base through *EpiVir* for itself and for its user. It should have fallacy rules in place and know how to apply tableau reasoning. It also needs a logical ontology created from the statements of its user (logical consequence). Grice's Maxim of Quality functions as an intentional probability operator to account for uncertainty about the user's statements. Tableau reasoning lets the robot decide if the human conclusions are true. Creative-proposition analysis tells the robot not to take an utterance literally. In its output, the robot may state the truthfulness it attaches to the user's conclusions but only if the user so requests. If the analyses of fallacies and tropes are correctly programmed (cf. Angleraud et al., 2019), the robot may also state which logical fallacy led to the wrongly stated conclusion or what rhetorical trope the user applied but by default, the robot does not do so (politeness rule).

8. Discussion/Conclusions

A robot should do what a user wants it to although what the user says is not always what s/he intends. Yet, the robot should stay polite and not return an error. We looked into the reasons why users make such mistakes and how they come to use figures of speech.

The root cause lies in the psychophysiological architecture of the brain (Crone & Konijn, 2018; Crone & Dahl, 2012; LeDoux, 1996/1999) in which all information runs through the limbic system, which basically responds with actions pertaining to joy (positive) and fear (negative). The neocortex has a more reflective and control function and may modulate the limbic responses through reasoning or solving problems creatively.

Whether in joy or fear but signal detection worsens when humans are emotionally excited (Cataldo & Cohen, 2015). In fear, people are biased to the detection of abnormality; they are quick to see cues to danger, giving rise to many 'false alarms' (Mujica-Parodi et al., 2017). We argued that if the neocortex responds to false positives through reasoning, *ex-consequentia* fallacies emerge. If approached creatively, the 'False positive' (*Fp*) type of metaphor occurs (e.g., 'Humans are robots'). With an aversion to robots, behavioral schemas and scripts reminiscent of machines are activated and attributed falsely to humans: We may call this phenomenon 'robotomorphism' (Jiménez-López, 2009, p. 80).

When in joy, people care less about detecting abnormality (cf. Bechara 2004, 2005); they miss out on cues to apparent robotic presence, leading to many 'false negatives.' We posited that from a reasoning viewpoint, dealing with false negatives produces *inverse errors*. Dealt with creatively, the 'False negative' (*Fn*) type of metaphor occurs ('My robot is my human partner'). In both cases, the normal human behavioral scripts are not deactivated, which then are attributed falsely to robots, resulting into anthropomorphism (e.g., Damiano & Dumouchel, 2018).

In an attempt to let a robot make sense of a user's logical fallacies and rhetorical tropes, we analyzed how to neutralize *ex-consequentia* reasoning and *inverse errors* and how to handle non-literal utterances such as metaphor and simile. For that, we needed four pieces of theory: Frege's distiction between reference and sense, Beth's tableau reasoning, Grice's communicative Maxim of Quality, and Hoorn's Epistemics of the Virtual. Frege taught us that a word may be without reference but not without meaning. Beth showed how to prove fallacy of a syllogism while Grice inspired us to formulate the MQ operator to bear with the user's illogical and non-literal statements. Epistemics of the Virtual made it possible to handle the fiction of possible worlds conjured up by illogical and non-literal communications.

These exercises let us formulate a fallacy fail-safe formula with epistemic and intensional operators as well as a creative-proposition analysis, packed together into a fail-safe protocol for fallacies and tropes, which may lead the way to a more polite handling by the robot of a user's something unintelligible remarks and commands. We asserted that the fail-safe formulas are hypotheses that in future research should be checked through corpus analytics and user tests.

We took a small step into a direction that usually evades logical analysis: Making sense of fallacies and analyzing figurative speech. We realize, however, how small the step is. With our protocol, the robot hopefully is able to handle clear-cut cases of illogical and non-literal utterances but in real life, things usually are more convoluted. Next, we provide two examples of fallacy and metaphor that are so embedded that a robot still cannot make sense of them, we suppose.

Hard to do: embedded fallacy

We will start from reasoning. What our fallacy fail-safe cannot do yet is to quarantine fallacies that are embedded in locally logic constituents that together are fallacious. Certainly, robots may be able to recognize *modus ponens* (i) and deem the logic form correct:

If A then B

A

Then B

or (ii):

If A then B
If B then C

A

Then C.

For example:

"The development of full artificial intelligence could spell the end of the human race." "It would take off on its own, and re-design itself at an ever increasing rate." "Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded." (Stephen Hawking in Cellan-Jones (2014))

Brought back to (ii):
If AI is autonomous (A) → it will redesign itself (B)
If redesigned (B) → humans will be superseded (C)
If superseded (C) → the human race will end (D)
If AI is autonomous (A) → the human race will be superseded (C) and thus end (D)

The fallacy outlined above has an embedded form. The constituents are logically sound but the overall reasoning is not. The elaborate *modus ponens* (ii) is used as a line of argumentation within a larger *ex-consequentia* structure:

If there is danger, I feel anxious
I feel anxious about AI (because A → B, B → C, C → D, A → D)
Therefore, AI must be dangerous

Counter examples against the fear of AI, for instance, may be that AI will not become completely autonomous because humans are not either, that humans may prevent it, that other AI may prevent it, that it is speculative that autonomy leads to self-redesign, that redesign does not necessarily lead to being superseded (perhaps there will be peaceful coexistence and collaboration), that being superseded will end the human race (perhaps we will merely serve but not become extinct, perhaps AI will protect us), and that most of these ideas may come from graphical Sci-Fi horror movies, not from deep knowledge about AI.

Hard to do: embedded metaphor

The next example is more creative and sketches a scenario in which metaphor is engrained in human conversation and has no simple A is B form. Suppose a little girl is on the plane with a cuddle robot in her arms. The robot has conversational AI and is used as a companion for the long flight and as a monitor for her safety. Right before take-off, the girl says: "Robot, I cannot put on my seat belt." Robot infers: If seat belt is not fastened, the girl's safety (goal) is at risk (negative outcome expectancies), she must fasten her seat belt. Then the girl explains: "I have no legs." Robot internally responds: False. My O_R tells me that this girl does have legs. The girl must be mistaken. At this point, the robot does not recognize the implicit metaphor because on the outer surface of the conversation, there is no category-mismatch; it is just the denial of a feature that manifestly is visible in the set.

We saw that in O_R, within the larger conception of Reality $\dot{R}$, entities that have sense rather than reference fall under the heading of Fiction $\dot{F}$ (Fig. 6, right). In Fiction $\dot{F}$, the girl's utterance should be taken as an *'as if'* statement and the robot should have had knowledge that someone without legs cannot put on a seatbelt. Under $\Diamond_{MQ}$, however, the nearest realistic interpretation to the girl's hyperbole is that she means she is too small for the seatbelt to fit, which is something the robot cannot guess.

Then the metaphor arises. The girl says: "I am a mermaid." The robot internally responds: False. My O_R tells me that this agency is a girl. Yet,

the metaphor now is manifest at the surface level. While running *EpiVir*, the robot keeps the fiction module $\dot{F}$ active and starts the creative-proposition analysis. It searches for females without legs but should limit its search to this child's knowledge base (O_h). Then the robot should find a fitting item, mermaid, and conclude: She probably means 'I am like a mermaid' because in fairy-tales $(O_h(\dot{F}))$ mermaids are girls with no legs, which she equals with her being too small in $(O_h(\dot{R}))$. The robot has searched the child's ontology to bring yet reference to the sign with sense alone (Fig. 1).

Hard to do: word error

What the mermaid-example tell us is that it is a general problem of Natural Language Processing (NLP) systems not to be able to discern between error and intended meaning because 'language understanding' (rather than 'processing') is still hard for any linguistic technology. Even humans have trouble discerning between a word error and a non-literal utterance such as a joke or poetic description, which people often explicitly mark as 'Just kidding' or ;-). Additionally, many artists welcome 'serendipitous findings,' so that the initial error becomes part of the artistic design.

As is, our fail-safe protocol will not distinguish between a word error and a non-literal utterance. When word errors are added to the ontology of the human (O_H), this may pose problems, because the ontology is what would be used by the language understanding module to correct that word.

One issue to tackle is mentioned by Postma, Ilievski, Vossen, and Van Erp (2016) and Ilievski, Postma, and Vossen (2016), remarking that most NLP systems opt for the high-frequency meaning of a word for which there is plentiful data available, whereas the peripheral meanings (e.g., connotations) are hardly detected. Vossen, Báez, Bajčetić, and Kraaijeveld (2018) and Vossen, Báez, Bajčetić, Bašić, and Kraaijeveld (2019) underscore that location-related object detection and the relevance of those objects to the conversation partners are important issues in language understanding that need to be resolved (but are not as of yet) for establishing a shared world and a base for communication. Thus, word errors should be addressed before our system can be effectively used but that holds for any other language-and-speech understanding system as well.

An initial solution is that the robot keeps an ontology for the user (O_H) and one for itself (O_R), so that the robot not necessarily goes astray by the misconceptions (whether intended or not) of its user. We cannot assume that any utterance (also between humans) is 100% correctly understood and so our system – like humans – has to deal with probabilities. The focus of the fail-safe protocol is on taking one step into the direction of solving the problem of logical fallacies and non-literal expressions, not so much on preprocessing the language before it can be used.

Hard to do: setting up ontologies

The fail-safe procedure also is not about setting up ontological databases. There are many ways an ontology can be constructed. It can be created by the system designer ('profile information') and updated by the user by hand. It could employ an AI that automatically analyzes language according to a machine-learning algorithm of choice. It could use Deep Learning services trained on Big Data. The Grounded Reference and Source Perspective of Fokkens, Vossen, Rospocher, Hoekstra, and Hage (2017), and Vossen et al. (2018) derives knowledge from different sources and stores it in RDF triples ready for reasoning. There are chit-chat systems under development that collect data and train models to update the profile information and personalize the dialogue, trying to predict the next utterance of the user (Zhang et al., 2018). In the Supplementary Materials, we do an exercise with a real-life example to explore the requirements on a fully functional language-understanding system that has our fail-safe protocol implemented.

With our fail-safe protocol for fallacies and tropes we may have contributed to better human-robot conversation but only for very clear and limited cases. Future research will focus on an implementation of the fail-safe protocol (see Supplementary Materials), running simulations, and conduct user tests to hopefully bring us to more advanced iterations of our groundwork of today.

8.1. Conclusions

We argued that in response to a robot:

- Reflection and affect are parallel brain-processes
- Negatively valenced action tendencies likely lead to false alarms; positively valenced tendencies to missed signals
- Negative tendencies regulated through reasoning likely provoke *ex-consequntia* fallacies; through creativity, they render *Fp-type* of metaphors and similes. The related fundamental attribution error is to assume machine-like qualities in humans (i.e. robotomorphism)
- Positive tendencies regulated through reasoning likely exert *inverse errors*; through creativity, they yield *Fn-type* of metaphors and similes. The fundamental attribution error is to assume human-like qualities in machines (i.e. anthropomorphism)

We proposed that:

- Conversational repair by robots in interaction with humans may come from a four-pronged approach: Ontologically, the robot should discern reference from sense, do tableau reasoning when detecting fallacy, apply the maxim of quality to 'pardon' its user for illogicality or non-literalness, and seek possible reference for sentence constituents that have sense alone (e.g., metaphors, simile)
- In future work, the resulting fail-safe protocols should be tested through corpus analytics, simulations, and user tests

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by *Communicating with and Relating to Social Robots* (NWO Open Competition Digitalisation, grant 406. DI.19.005). The authors wish to thank Jan Treur and Qixin Wang for reviewing the logics. We are grateful to Rick Cooper for commenting on an earlier draft of this paper. We kindly acknowledge the anonymous reviewers for their valuable comments and suggestions. Ivy S. Huang is thanked for her translation of the abstract into Chinese. The authors have no competing interests to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogsys.2021.10.003.

References

Angleraud, A., Houbre, Q., & Pieters, R. (2019). Teaching semantics and skills for human-robot collaboration. *Paladyn, Journal of Behavioral Robotics, 10*(1), 318–329. https://doi.org/10.1515/pjbr-2019-0025

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics, 7*(4), 465–478. https://doi.org/10.1007/s12369-015-0298-7

Bartneck, C., Reichenbach, J., & Carpenter, J. (2006). Use of praise and punishment in human-robot collaborative teams. *Proceedings of the 15th IEEE International Symposium 2006, Robot and Human Interactive Communication (RO-MAN '06) Sept. 6-8, 2006. Hatfield, Herthfordshire, UK* (pp. 177-182). Piscataway, NJ: IEEE. doi: 10.1109/ROMAN.2006.314414.

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition, 55*(1), 30–40.
Bechara, A. (2005). Decision making: Impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience, 8*(11), 1458–1463.
Bernsen, N. O., Dybkjaer, H., & Dybkjaer, L. (2014). *Designing interactive speech systems: From first ideas to user testing.* London: Springer.
Beth, E. W. (1955). *Semantic entailment and formal derivability*. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.
Beth, E. W. (1956). *A semantic construction of intuitionistic logic*. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.
Blanchette, I., & Richards, A. (2004). Reasoning about emotional and neutral materials: Is logic affected by emotion? *Psychological Science, 15*(11), 745–752. https://doi.org/10.1111/j.0956-7976.2004.00751.x
Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*(1), 627–652.
Caporael, L. R. (2006). Three tips from a social psychologist for building a social robot. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control (AMC '06) March 27-29, 2006. Istanbul, Turkey* (pp. 744-749). Piscataway, NJ: IEEE. doi: 10.1109/AMC.2006.1631753.
Cataldo, A. M., & Cohen, A. L. (2015). The effect of emotional state on visual detection: A signal detection analysis. *Emotion, 15*(6), 846–853. https://doi.org/10.1037/emo0000091
Cellan-Jones, R. (2014, January 2). Stephen Hawking warns artificial intelligence could end mankind. *BBC News.* Retrieved January 30, 2018 from https://www.bbc.com/news/technology-30290540.
Ceschi, A., Costantini, A., Sartori, R., Weller, J., & Di Fabio, A. (2019). Dimensions of decision-making: An evidence-based classification of heuristics and biases. *Personality and Individual Differences, 146*, 188–200. https://doi.org/10.1016/j.paid.2018.07.033
Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nature Reviews Neuroscience, 13*(9), 636–650. https://doi.org/10.1038/nrn3313
Crone, E. A., & Konijn, E. A. (2018). Media use and brain development during adolescence. *Nature Communications, 9*(588), 1–10. https://doi.org/10.1038/s41467-018-03126-x
Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human-robot co-evolution. *Frontiers in Psychology, 9*, 468. https://doi.org/10.3389/fpsyg.2018.00468
Darwin, C. (1872/2002). *The expression of emotions in man and animals.* New York: Oxford University.
Davis, M. A. (2009). Understanding the relationship between mood and creativity: A meta-analysis. *Organizational Behavior and Human Decision Processes, 108*(1), 25–38. https://doi.org/10.1016/j.obhdp.2008.04.001
Diano, M., Tamietto, M., Celeghin, A., Weiskrantz, L., Tatu, M.-K., Bagnis, A., … Costa, T. (2017). Dynamic changes in amygdala psychophysiological connectivity reveal distinct neural networks for facial expressions of basic emotions. *Nature Scientific Reports, 7*(1). https://doi.org/10.1038/srep45260
Dolan, R. J. (2007). The human amygdala and orbital prefrontal cortex in behavioural regulation. *Philosophical Transactions of the Royal Society B, Biological Sciences, 362*(1481), 787–799. https://doi.org/10.1098/rstb.2007.2088
Domes, G., & Zimmer, P. (2019). Acute stress enhances the sensitivity for facial emotions: A signal detection approach. *Stress, 22*(4), 455–460. https://doi.org/10.1080/10253890.2019.1593366
Dreisbach, G. (2006). How positive affect modulates cognitive control: The costs and benefits of reduced maintenance capability. *Brain and Cognition, 60*(1), 11–19.
Dreisbach, G., & Goschke, T. (2004). How positive affect modulates cognitive control: Reduced perseveration at the cost of increase distractibility. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30*, 343–353.
Engelhard, I. M., & Arntz, A. (2005). The fallacy of ex-consequentia reasoning and the persistence of PTSD. *Journal of Behavior Therapy and Experimental Psychiatry, 36*(1), 35–42. https://doi.org/10.1016/j.jbtep.2004.11.004
Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition, 26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psycholical Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978
Faust, M. (2012). Thinking outside the left box: The role of the right hemisphere in novel metaphor comprehension. In M. Faust (Ed.), *The handbook of the neuropsychology of language* (pp. 425–448). Chichester, UK: Blackwell.
Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., & Hage, W. (2017). GRaSP: Grounded Representation and Source Perspective. In *Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities (RANLP'17). Varna, Bulgaria* (pp. 19-25). doi: 10.26615/978-954-452-040-3_003.
Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology, 2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300
Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56*(3), 218–226. https://doi.org/10.1037/0003-066X.56.3.218
Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion, 19*(3), 313–332. https://doi.org/10.1080/02699930441000238
Frege, G. (1892). Ueber Sinn und Bedeutung [On sense and reference]. *Zeitschrift für Philosophie und philosophische Kritik, 100*, 25–50.
Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic.
Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University.
Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010). Critic, compatriot, or chump?: responses to robot blame attribution. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot interaction (HRI '10) March 02-05, 2010. Osaka, Japan* (pp. 211-218). Piscataway, NJ: IEEE.
Haring, K. S., Watanabe, K., Velonaki, M., Tossell, C. C., & Finomore, V. (2018). FFAB - the Form Function Attribution Bias in human-robot interaction. *IEEE Transactions on Cognitive and Developmental Systems, 10*(4), 843–851. https://doi.org/10.1109/TCDS.2018.2851569
Hauser, D. J., Nesse, R. M., & Schwarz, N. (2017). In *The Science of Lay Theories* (pp. 341–354). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57306-9_14.
Hildebrandt, C., Törsleff, S., Caesar, B., & Fay, A. (2018). Ontology building for cyber-physical systems: a domain expert-centric approach. In *Proceedings of the 14th IEEE International Conference on Automation Science and Engineering (CASE '18) Aug. 20-24. Munich, Germany* (pp. 1079-1086). Piscataway, NJ: IEEE. doi: 10.1109/COASE.2018.8560465.
Hoffman, R. R., Cochran, E. L., & Nead, J. J. (1994). Cognitive metaphors in experimental psychology. In D. E. Leary (Ed.), *Metaphors in the history of psychology* (pp. 173–229). New York: Cambridge University.
Hoorn, J. F. (2012). *Epistemics of the virtual*. Philadelphia, PA: John Benjamins.
Hoorn, J. F., Konijn, E. A., & Pontier, M. A. (2018). Dating a synthetic character is like dating a man. *International Journal of Social Robotics, 1–19*. https://doi.org/10.1007/s12369-018-0496-1. Available from https://link.springer.com/article/10.1007/s12369-018-0496-1
Humphrey, M. K., Bryson, F. M., & Grimshaw, G. M. (2010). Metaphor processing in high and low schizotypal individuals. *Psychiatry Research, 178*(2), 290–294.
Huang, Y., Xie, J., & Wang, R. (2018). Electroencephalography delta, theta, and alpha oscillations in valence-space metaphorical associations. *NeuroReport, 29*(12), 1017–1022.
Hung, E. C. K. (2016). Tackling design fixation of cultural product designers through homeomorphism. In W. Chung, C. Shin (Eds.), *Advances in affective and pleasurable design, v. 483* (pp. 491-498). Cham, CH: Springer. doi.org/10.1007/978-3-319-41661-8_47.
Ilievski, F., Postma, M., & Vossen, P. (2016). Semantic overfitting: What 'world' do we consider when evaluating disambiguation of text? In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16): Technical Papers, Dec. 11-17, 2016. Osaka, Japan.* (pp. 1180-1191). COLING 2016 Organizing Committee.
Imanishi, K. (1941/2002). *A Japanese view of nature: the world of living things* [Seibutsu no Sekai]. P. J. Asquith (Ed.), P. J. Asquith, H. Kawakatsu, S. Yagi, & H. Takasaki (Trans.). New York: RoutledgeCurzon.
Isen, A. M., Johnson, M. M. S., Mertz, E., & Robinson, G. F. (1985). The influence of positive affect on the unusualness of word associations. *Journal of Personality and Social Psychology, 48*(6), 1413–1426. https://doi.org/10.1037/0022-3514.48.6.1413
Jwalapuram, P. (2017). Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop Associated (RANLP'17) Sept. 2017. Varna, Bulgaria* (pp. 17-24). Moscow, Russia : INCOMA. doi: 10.26615/issn.1314-9156.2017_003.
Jiménez-López, E. (2009). General systems Weltanschauung. In F. Parra-Luna (Ed.), *System science and cybernetics, II. Encyclopedia of life support systems* (pp. 59–81). Oxford, UK: EOLSS.
Juhasz, J. B., & Sarbin, T. R. (1966). On the false alarm metaphor in psychophysics. *The Psychological Record, 16*(3), 323–327.
Konijn, E. A., & Hoorn, J. F. (2016). Empathy with and projecting feelings onto robots from schemas about humans. *International Journal of Psychology, 51*(S1), 837.
Konijn, E. A., & Hoorn, J. F. (2017). Parasocial interaction and beyond: Media personae and affective bonding. In P. Roessler, C. Hoffner, & L. van Zoonen (Eds.), *The international encyclopedia of media effects* (pp. 1–15). NY: Wiley-Blackwell. https://doi.org/10.1002/9781118783764.wbieme0071.
Konijn, E. A., van der Molen, J. H. W., & van Nes, S. (2009). Emotions bias perceptions of realism in audiovisual media: Why we may take fiction for real. *Discourse Processes, 46*(4), 309–340. https://doi.org/10.1080/01638530902728546
Lakoff, G., & Johnson, M. (2003). *Metaphors we live by*. London: University of Chicago.
LeDoux, J. (1996/1999). *The emotional brain: The mysterious underpinnings of emotional life.* London: Weidenfeld & Nicolson.
Lepore, E., & Stone, M. (2015). *Imagination and convention: Distinguishing grammar and inference in language*. Oxford, UK: Oxford University.
Lerche, V., Christmann, U., & Voss, A. (2018). Impact of context information on metaphor elaboration. *A diffusion model study. Experimental Psychology, 65*(6), 370–384. https://doi.org/10.1027/1618-3169/a000422
Magnani, L. (2015). Naturalizing logic: Errors of reasoning vindicated: Logic reapproaches cognitive science. *Journal of Applied Logic, 13*(1), 13–36. https://doi.org/10.1016/j.jal.2014.11.001
Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology, 13*(1–2), 125–137. https://doi.org/10.1207/S15327663JCP13-1&2_11
Mujica-Parodi, L. R., Cha, J., & Gao, J. (2017). From anxious to reckless: A control systems approach unifies prefrontal-limbic regulation across the spectrum of threat detection. *Frontiers in Systems Neuroscience, 11*, 1–18. https://doi.org/10.3389/fnsys.2017.00018
Murray, N., Sujan, H., Hirt, E. R., & Sujan, M. (1990). The influence of mood on categorization: A cognitive flexibility interpretation. *Journal of Personality and Social Psychology, 59*(3), 411–425. https://doi.org/10.1037/0022-3514.59.3.411

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Nygren, T. E., Isen, A. M., Taylor, P. J., & Dulin, J. (1996). The influence of positive affect on the decision rule in risk situations: Focus on outcome (and especially avoidance of loss) rather than probability. *Organizational Behavior Human Decision Processes, 66* (1), 59–72.

Pfenniger, K. H. (2001). The evolving brain. In K. H. Pfenniger, & V. R. Shubik (Eds.), *The origins of creativity* (pp. 89–97). New York: Oxford University.

Postma, M., Ilievski, F., Vossen, P., & Van Erp, M. (2016). Moving away from semantic overfitting in disambiguation datasets. In A. Louis, M. Roth, B. Webber, M. White, & L. Zettlemoyer (Eds.), *Proceedings of the EMNLP Workshop on Uphill Battles in Language Processing. Nov. 2016. Austin, TX* (pp. 17-21). Association for Computational Linguistics. doi: 10.18653/v1/W16-6004.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology, 10* (pp. 173–220). New York: Academic.

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: My research journey. *Perspectives on Psychological Science, 13*(6), 750–769. https://doi.org/10.1177/1745691618769855

Shutova, E., Sun, L., Gutiérrez, E. D., Lichtenstein, P., & Narayanan, S. (2017). Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics, 43*(1), 71–123.

Sorower, M. S., Doppa, J. R., Orr, W., Tadepalli, P., Dietterich, T. G., & Fern, X. Z. (2011). Inverting Grice's maxims to learn rules from natural language extractions. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11) Dec. 12 - 15. Granada, Spain* (pp. 1053-1061). Red Hook, NY: Curran Associates.

Thicknesse, P. (1784). *The speaking figure, and the automaton chess player, exposed and detected.* London: John Stockdale.

Thompson, N. M., Uusberg, A., Gross, J. J., & Chakrabarti, B. (2019). Empathy and emotion regulation: An integrative account. In N. Srinivasan (Ed.), *Progress in brain research, 247* pp. 273–304). Cambridge, MA: Academic. https://doi.org/10.1016/bs.pbr.2019.03.024.

Von Bertalanffy, L. (1968). *General system theory. Foundations, development, applications.* New York: George Braziller. Available from https://monoskop.org/images/7/77/Von_Bertalanffy_Ludwig_General_System_Theory_1968.pdf.

Vossen, P., Báez, S., Bajčetić, L., Bašić, S., & Kraaijeveld, B., (2019). A communicative robot to learn about us and the world. In V. Selegey (Ed.), *Proceedings the International Conference (Dialogue'19), May 29 - June 1, 2019. Moscow, RSHU* (vol. *18* (25)). Computer Linguistics and Intelligent Technologies. Available from http://www.dialog-21.ru/media/4636/vossenpplusetal-050.pdf.

Vossen, P., Báez, S., Bajčetić, L., & Kraaijeveld, B. (2018). Leolani: a reference machine with a theory of mind for social communication. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *International Conference on Text, Speech, and Dialogue (TSD'18), Sept. 11. Brno, Czech Republic. Lecture Notes in Computer Science, 11107* (pp. 15-25). Cham, CH: Springer. doi: 10.1007/978-3-030-00794-2_2.

Xie, W., & Zhang, W. (2014). Contributions of cognitive factors in conceptual metaphors. *Metaphor and Symbol, 29*(3), 171–184. https://doi.org/10.1080/10926488.2014.924282

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv*: 1801.07243.

Złotowski, J. A., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics, 7*(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6

Further reading

Swaab, D. (2016). *Ons creatieve brein. Hoe mens en wereld elkaar maken* [Our creative brain. How humans and world make each other]. Amsterdam: Atlas Contact.