2. Disciplinary Context and Terminology

Before we continue our tour along the life cycle of robots and explore how in/animacy is attributed to robot technology in different contexts, we first need to equip ourselves with some conceptual tools.

The question of when and why humans attribute characteristics of living beings to non-living entities, or characteristics of humans to non-human entities, has been a topic of interest for several scientific fields. Within and across these different disciplines a range of terminology is employed to describe the same or similar phenomena. The present chapter will untangle this complex disciplinary, historical, and terminological context.

First, the chapter will show how human-robot interaction (HRI) research approaches the phenomenon of animacy attribution to robots. We will explore the field's strongly innovation- and application-driven approach towards the phenomenon, and explore basic assumptions underlying this research, as well as methodological and ethical issues discussed in this context.

Second, we will take apart the tangle of different terms used across disciplines – such as "anthropomorphism", "animacy", "intentionality", and "agency" – and establish the use of the term "attribution of animacy" for the purpose of this book.

Third, the chapter will give an overview of further relevant disciplinary perspectives on the topic. It will show that, historically, phenomena like anthropomorphism have often been viewed either as a "primitive" interpretation of environmental cues or, in the context of academia, as methodological malpractice. Only relatively recently has the topic drawn scientific interest as an object of research in itself. We will see how different disciplines approach issues like anthropomorphism, animacy detection, and technological agency.

2.1. Human-Robot-Interaction Research: "Controlling" In/Animacy Attributions

In Chapter 1 (Section 1.1), we already touched upon the fact that most academic attention on the phenomenon of animacy attributions to robots can be observed in the context of human-robot interaction (HRI) research. Most HRI research takes place in the field of social robotics, focusing on robots explicitly meant for more complex user interactions, usually with a humanoid design. These robots can be either bespoke platforms or off-the-shelf models like Softbank's Nao[1], which are still too expensive for the average customer. Robot technology safe and robust enough to be employed in direct physical contact with humans is only just now becoming available and affordable enough for the mass market, for example in the form of lightweight robot arms, household robots, or small tele-operated platforms. These robots are usually not intended for complex social interactions. Nonetheless, there is a slowly growing awareness of the complexity of interaction with "mechanical looking" and seemingly "non-social" robots not specifically designed to interact with humans or to appear life-like in any way. Andrea Guzman (2016), for example, argues for the designation of industrial and manufacturing machines as technologies of communication.

The goal of research efforts in HRI is usually not the short-term realization of an interaction scenario representing the current state of robot technology. Instead, HRI research usually focuses on the exploration of scenarios expected to become relevant only in the future, such as the coexistence with very human-like robots (Bischof, 2015, p. 211). In order to simulate the anticipated capabilities of future robots some interaction studies make us of so-called "Wizard-of-Oz" experiments, in which the robot's behavior is secretly controlled by a human operator. This method is sometimes criticized – both for its deception of study participants and for not really studying human interaction with the robot, but rather with the human robot operator, only relayed through a robot (cf. Laurel D. Riek, 2012).

For most people, the long-term use of robot technology, especially socially interactive robot technology, is not yet an everyday practice. Most HRI research therefore studies short-term interactions in laboratory settings, or explores attitudes towards robots based on people's existing knowledge and imagination of robots.

1 https://www.softbankrobotics.com/emea/en/nao (accessed 2019-12-21).

Notable exceptions to this approach are studies exploring some of the few contexts where users already closely interact with robots every day. This includes the professional use of remote controlled robot platforms for explosive ordnance disposal (J. Carpenter, 2013, 2016) or search and rescue efforts (Bethel & Murphy, 2006; Murphy, Riddle, & Rasmussen, 2004), and the use of vacuum cleaner robots in private households (Forlizzi & DiSalvo, 2006; Sung et al., 2007). These field studies, which all focus on non-humanoid robots, only superficially deal with the question of how or why humans attribute characteristics of living beings to robots. In the more short-term, laboratory-based research, however, this topic receives plenty of scholarly attention. Especially in the field of social robotics, a range of studies investigates what is sometimes called "anthropomorphic projections" or "anthropomorphic attributions"[2] to robots (Damiano & Dumouchel, 2018, p. 2) – the perception of robots as having human-like characteristics. This research is frequently based on the findings of cognitive and evolutionary psychology – an aspect explored further in Section 2.3 of this chapter. The methodological spectrum and quality of these studies is broad. It ranges from psychophysiological and neuroscience methods, to behavioral observation, to cognitive tests and self-assessment questionnaires.[3] While the majority of these studies uses "homemade" methodological tools, there are some efforts to standardize the "measurement" of anthropomorphic projections, such as the Godspeed questionnaire (Bartneck et al., 2009; Ho & MacDorman, 2010) or the RoSAS scale (Carpinella et al., 2017).

Underlying this approach is an idea inherent to the innovation-driven interests and methods of HRI research: that of "controlling" humans' anthropomorphic attributions to technology. In contrast to the historical skepticism towards anthropomorphism and similar phenomena in other academic disciplines (which we will explore in Section 2.3), the self-imposed challenge for HRI research is "not how to avoid anthropomorphism, but rather how to embrace it" (Duffy, 2003a, p. 180). The goal is to identify user and robot characteristics involved in the "activation" of anthropomorphic attributions. These characteristics are then supposed to act as predictors for specific behavioral

2 Section 2.2 will explain why this book uses a different term for the same phenomenon ("animacy attribution").

3 For a general overview see e.g. Złotowski et al. (2018) or Damiano and Dumouchel (2018). For an overview of neuroscience approaches to human–robot interaction research see Henschel, Hortensius & Cross (2020).

and emotional reactions of the user, which in turn are understood as indicators for the "strength" of the anthropomorphic attribution.

Typical user characteristics used as independent variables in this type of HRI research are standard demographic variables such as age, gender, or cultural background, as well as personality traits.[4] The catalog of variables also includes complex (and difficult to operationalize) traits such as "loneliness", "need for control", "experience with robots" and "interest in technology".[5] On the side of the robot, HRI research explores relatively simple variables like size, color, or material, but also more complex, usually unstandardized factors such as "physical presence", "human likeness", "animacy", or "behavioral complexity".[6]

Combinations of these independent variables are then explored in their effect on various emotional and behavioral measures. Some of these are meant to quantify the "amount" of human-likeness study participants attribute to robots. Here, we find studies observing "intelligence attribution", "mind perception", "perceived social presence", "perceived sociability", but also "embarrassment from being observed by the robot", "empathy with the robot" and "hesitation to switch off", or refusal to physically "harm", or "kill" a robot.[7]

---

4 Selected examples: Age (e.g. Kuo et al., 2009; Reich & Eyssel, 2013); gender (e.g. Chin, Sims, Clark, & Lopez, 2004; De Graaf & Ben Allouch, 2013; T. Nomura, Kanda, Suzuki, & Kato, 2008; Schermerhorn, Scheutz, & Crowell, 2008); cultural background (e.g. Evers, Maldonado, Brodecki, & Hinds, 2008; Tatsuya Nomura et al., 2008); personality traits (e.g. Syrdal, Dautenhahn, Woods, Walters, & Kheng Lee Koay, 2006; Walters et al., 2005; Woods et al., 2007).

5 Selected examples: Loneliness (e.g. Epley, Akalis, Waytz, & Cacioppo, 2008; Reich & Eyssel, 2013); need for control (e.g. Epley, Waytz, Akalis, & Cacioppo, 2008); interest in/experience with robots or technology (e.g. Bartneck, Suzuki, Kanda, & Nomura, 2007; European Commission, 2012; Heerink, 2011, 2011; Tatsuya Nomura, Suzuki, Kanda, Yamada, & Kato, 2011; Reich & Eyssel, 2013; Woods et al., 2007).

6 Selected examples: Size (e.g. Walters, Koay, Syrdal, Campbell, & Dautenhahn, 2013); color/material (e.g. J. Wright, Sanders, & Hancock, 2013); physical presence (e.g. Kidd & Breazeal, 2004); human likeness (e.g. Bartneck, Bleeker, Bun, Fens, & Riet, 2010; L. U. Ellis et al., 2005; Hinds, Roberts, & Jones, 2004; Kiesler & Goetz, 2002; R. H. Kim, Moon, Choi, & Kwak, 2014; Kwak, 2014; von der Pütten & Kramer, 2012); animacy (e.g. Bartneck, Kanda, Mubin, & Al Mahmud, 2009); behavioral complexity (e.g. Rau, Li, & Liu, 2013; Scholl & Tremoulet, 2000; Vouloutsi, Grechuta, Lallée, & Verschure, 2014).

7 Selected examples: Intelligence attribution (e.g. H. M. Gray, Gray, & Wegner, 2007; Kiesler & Goetz, 2002; Sung, Guo, Grinter, & Christensen, 2007); mind perception/mind attribution (e.g. Epley, Waytz, & Cacioppo, 2007; H. M. Gray et al., 2007; Kamide,

Studies often try to explore the effect of different variables on the "success" of human-robot interaction by measuring participants' "attitude towards robots", "willingness to use robots", or "acceptance of robots".[8] In "an attempt to design and control not only robotics systems but also the entire process of human-robot interaction, users' performance included" (Zawieska, 2015, p. 3) these insights are meant to help with the improvement of future human-robot-interaction. Not surprisingly for such a complex issue, and considering the jumble of different variables, few widely accepted theories or models have emerged so far. One exception is Nicholas Epley and colleagues' (2007) Three-Factor Theory of Anthropomorphism, which strives to integrate the various perspectives investigated in previous studies, and suggests three "psychological determinants" of anthropomorphism: the accessibility and applicability of anthropocentric knowledge, the motivation to explain and understand the behavior of other agents, and the desire for social contact and affiliation. While Epley and colleagues explicitly named robotics as an area of application, and the model is frequently referenced in the HRI literature, they are not HRI researchers, but cognitive scientists. The success of their model in the HRI community demonstrates the close connection of HRI and the cognitive sciences in this particular context (which Section 2.3 will explore in depth).

Two basic assumptions underlie many of the HRI studies trying to find predictors for anthropomorphic attributions to robots: Firstly, the assumption that it is possible to "switch on" anthropomorphic attributions with the right kind of robot design or robot behavior. Secondly, the assumption that anthropomorphic attributions to robots are desirable and advantageous for human-robot-interaction.

Eyssel, & Arai, 2013); perceived social presence or sociability (e.g. Choi, Kim, & Kwak, 2014; Kiesler & Goetz, 2002; R. H. Kim et al., 2014; Schermerhorn et al., 2008); embarrassment (e.g. Bartneck et al., 2010; Choi et al., 2014); empathy (e.g. Darling, Nandy, & Breazeal, 2015; Riek, Rabinowitch, Chakrabarti, & Robinson, 2009a, 2009b; A. M. Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013); hesitation to switch off/harm/"kill" robot (e.g. Bartneck, van der Hoek, Mubin, & Al Mahmud, 2007; Riek et al., 2009b, 2009a; Darling, 2012; A. M. Rosenthal-von der Pütten et al., 2013; Darling et al., 2015;).

8 Selected examples: Attitude towards/willingness to use/acceptance of robots (e.g. De Graaf & Ben Allouch, 2013; Kwak, Kim, & Choi, 2014; T. Nomura et al., 2008; T. Nomura, Kanda, Suzuki, Yamada, & Kato, 2009; Stafford, MacDonald, Jayawardena, Wegner, & Broadbent, 2014).

The first assumption is based on the notion that certain characteristics of robot technology – such as embodiment, mobility, autonomous behavior, or humanoid design – trigger a human perception system highly primed to recognize animacy (cf. Section 2.3). Commonly, "the robot is an inanimate object" is understood to be the default interpretation or null hypothesis. HRI studies then try to trigger anthropomorphic attributions in a controlled way, by manipulating the design or behavior of the robot. The idea behind this approach is that the existence or magnitude of certain features is able to push the robot over a "social threshold", giving it a "social presence" (Damiano & Dumouchel, 2018; cf. Levillain & Zibetti, 2017).

Few studies explicitly look at the opposite, at attributions of inanimacy. Presumably, it being considered the null hypothesis, this attributive perspective is not viewed as an interesting phenomenon per se. An exception to this is a cluster of research working with the concept of "dehumanization". Here, the idea is that "looking at a process of depriving objectified humans of characteristics regarded as crucial in order to be perceived and treated as a human" would contribute to "identify[ing] the key characteristics for robots to affect their anthropomorphism" (Złotowski et al., 2018, pp. 1 & 2; Waytz, Epley, & Cacioppo, 2010; Morera et al., 2018; cf. Haslam, 2006).

There are also some field studies finding anecdotal evidence of what we will also observe in the context of this book: robots being simultaneously enacted as an agent and as a thing, as both animate and inanimate. A study exploring spontaneous interactions with a social robot in a classroom setting found that for their study participants "seemingly contradictory features – a thing and a living creature – unproblematically coexist[ed]", "the robot present[ing] its multiple facets so that each theme c[ould] resurface at any particular moment" (Alač, 2016, pp. 12 & 15). A short field study with a hospital delivery robot found that the hospital staff perceived the robot as both a machine and a colleague, both "perspectives mutually coexist[ing], even for the same person" (Ljungblad et al., 2012, p. 9).

Studies like these, which consider attributions of both animacy and inanimacy, are rare, however. This might in part be due to the second assumption underlying many HRI studies: That of anthropomorphic attributions being beneficial for smooth human-robot interaction. They are thought to "facilitate … human-machine interaction, … increase people's willingness to care about the well-being of robots" (Złotowski et al., 2015, p. 351), and even to "facilitate … the introduction of robots in the society at large" (Ferrari, 2015, p. 17). This idea can also be encountered outside of a purely academic context.

For example, in the marketing campaign for the personal service robot Jibo, roboticist Cynthia Breazeal argued that "it is really important for technology to be humanized" (cited in Markoff, 2014). This assumption inspires HRI research efforts with an openly communicated agenda: In HRI, anthropomorphism is mainly studied in order to use the insights for the improvement of future robots' interaction capabilities and "usefulness … by creating social bonds that increase a sense of social connection" (Epley et al., 2007, p. 897). In this context, knowledge about anthropomorphism is now "highly valued by many roboticists and computer scientists" for its potential to be used as a means to control user reactions to robots (Vidal, 2007, p. 3). Research efforts exploring the processes behind anthropomorphism are therefore frequently fueled by the inherent goal of building socially interactive robots:

> "While anthropomorphism is clearly a very complex notion, it intuitively provides us with very powerful physical and social features that will no doubt be implemented to a greater extent in social robotics research in the near future." (Duffy, 2002, p. 5)

In some parts of the robotics community, building a "perfect" human-like robot is considered the ultimate goal, or even "holy grail" (e.g. Duffy, 2006, p. 33): "It seems a truth universally acknowledged that a roboticist with a good research lab must want to create a humanoid!"[9] (Keay, 2011, p. 66). This goal is also fueled by the tempting engineering challenge it poses. As one of the roboticists interviewed for this book (cf. Chapter 3) explained:

> "[A humanoid robot] is not only the most extreme form [of robot]; it is the most difficult form of autonomous systems you can work on. … It's the interactions, the possibilities of interaction … they basically explode." (R2.3-00:07:22-8)[10]

The challenge is not only to make a robot look like a human but, even more, to make it behave like a human. While it is relatively easy to put a realistic looking "skin" on a robot "skeleton", making a robot autonomously move and speak in a completely natural-appearing way is still an unsolved problem.

---

9 A reference to the famous first sentence of Jane Austen's "Pride and Prejudice" (1813): "It is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife."

10 The numbers after this quote refer to the position in the audio transcript of the interview.

Robots with relatively good interaction abilities are usually remote controlled or follow a predetermined script, rather than acting autonomously – such as Hanson Robotics' Sophia[11], or Hiroshi Ishiguro's various Geminoids[12]. However, a realistic humanoid design is not a prerequisite for humans to experience a strong social connection and to attribute emotions, desires, and even personality traits to robot technology (cf. Chapter 1).[13]

The strongly innovation-driven goal of many HRI projects – to find out which robot characteristics have to meet which human characteristics in order to attain the best possible "interaction experience" – faces criticism from both within and outside the HRI community. Not only is there a range of methodological problems, such as the questionable operationalization of complex concepts like "human-likeness". There is also no clear evidence that human likeness actually has a positive influence on human-robot interaction.

HRI researchers face a methodological challenge: On the one hand, studies exploring human-robot interaction are supposed to use "realistic" scenarios in order to make the results generalizable and maybe even usable for marketable applications. On the other hand, variables like robot and environment features or user reactions need to be measurable and comparable. The result is often a methodological compromise, with research being conducted in laboratory environments with simulated "real life" scenarios, and metrics constructed around what is doable within the constrictions of the institutional conditions and available resources (Meister, 2014, p. 120). Often, this results in the use of "i-method"-approaches[14], as well as a naïve and uncritical use of what is understood to be "social science" methods by untrained engineers (Irfan et al., 2018). This leads to a lack of common metrics, methods, and generalizability – making the findings of most HRI studies neither comparable to each other, nor generalizable to a real life environment or a wider population (Steinfeld et al., 2006; Dautenhahn, 2007; Bethel & Murphy, 2010; Bischof, 2015).

These operationalization issues are also present in research on anthropomorphism and related phenomena in the context of robotics. For example, there is no consensus on what "human-like" robot design or robot behavior

11 https://www.hansonrobotics.com/sophia (accessed 2019-11-19).

12 https://eng.irl.sys.es.osaka-u.ac.jp/robot (accessed 2019-11-19).

13 E.g. J. Carpenter, 2016; Julia Fink, 2014; Forlizzi & DiSalvo, 2006; Kolb, 2012; Levillain & Zibetti, 2017; Sandry, 2015b; Sung et al., 2007; Yang et al., 2015.

14 Implying a designer's reliance on personal experience, attempting to take on a layperson's perspective (cf. Oudshoorn, Rommes, & Stienstra, 2004).

means, or how to "measure" anthropomorphism and users' attributions of human-likeness to robots – making study results difficult or even impossible to compare. Those difficulties are observable in arbitrary categorizations of robot designs in many studies trying to contrast "humanoid" against "non-humanoid" robots, with every study drawing the border between the categories somewhere else. There have been proposals for universal categorizations, such as Brian Duffy's (2003b) "Anthropomorphism Design Space for Robot Heads", but most HRI studies use "homemade" categories. Sometimes these border on the absurd, like when a robot vacuum cleaner with googly eye stickers is categorized as "human oriented"[15] (Kwak, 2014), or when an oven with arms is supposed to be "anthropomorphic" (Osawa, Mukai, & Imai, 2007). For one end of the design spectrum, there is at least a term most agree on – "humanoid". For the other end, a plethora of terms is in use, including "non-humanoid", "mechanistic", "mechanoid", "mechanical", "appearance-constrained", "single purpose", "functional", and "with few anthropomorphic features".

There is also no generally accepted measure for the "strength" of users' anthropomorphic attributions to robots. Most HRI studies do not operationalize this at all, but instead directly investigate the influence of different "human-likeness" levels on users' attributions of "mind", "sociability", or "intimacy", or emotional and behavioral reactions like empathy (e.g. Carpenter 2013; Garreau 2007; Garber 2013; Riek et al. 2009), embarrassment (e.g. Choi et al. 2014; Bartneck 2010), or decision making (e.g. Bartneck et al. 2007; Chandler & Schwarz 2010).

The overall methodological disunity is a topic of discussion within the HRI community. There are efforts for finding some consensus and comparability, for example by trying to make anthropomorphism "measurable" on a one- or two-dimensional scale (Bethel & Murphy, 2010; Ruijten et al., 2019), or by developing standardized tests for anthropomorphic attributions to robots (e.g. Bartneck et al., 2009; cf. Murphy & Schreckenghost, 2013). However, no generally accepted approach has been agreed on yet.

HRI researchers not only disagree on how to measure a robot's human-likeness and users' reactions to it. There is also no consensus on whether making a robot human-like is actually desirable. While there has been a steady stream of social robotics research based on the assumption that giving a robot the "right set" of lifelike features will somehow make users able and willing to

15 The vacuum cleaner without googly eyes, meanwhile, was categorized as "product oriented" (Kwak, 2014).

interact with the robot, and the construction of a "realistic" humanoid robot is considered by some as the "holy grail" of robotics, there is actually no consent within the robotics and HRI community on whether making a robot as humanlike as possible is worthwhile. In the context of this discussion, the so-called "Uncanny Valley" effect is referenced frequently. First proposed by Japanese roboticist Masahiro Mori (1970), the concept hypothesizes that the relation of a robot's human-likeness and observers' emotional responses is not linear. The underlying idea is that the more a robot is designed to look and behave like a human, the more positive observers react to it. With one crucial exception: If a certain level of human-likeness is reached – the robot resembling a real human very closely, but falling short of being a perfect representation – observers' reactions are adverse, even disgusted. The sharp dip in the graph representing the relationship of human-likeness and observer reactions is referred to as the Uncanny Valley. Although there is no clear empirical support for the hypothesis, even after decades of research, it is referenced frequently in the HRI literature (Brenton et al., 2005; Bartneck et al., 2009; Damiano & Dumouchel, 2018; MacDorman, 2019). However, the discussion surrounding the validity of the Uncanny Valley concept does direct attention to one important issue, namely that of the expectations a human-like robot raises with users. One of several explanations put forward for the (presumed) Uncanny Valley effect is that a very human-like robot design causes human users to have certain expectations about the robot's behavior, such as realistic movements or a smooth natural language interaction. At the current state of technology however, no humanoid robot is able to fulfill these expectations to a satisfactory level and – so the idea – the ensuing disappointment, irritation, or even disgust experienced by the user causes the Uncanny Valley effect (Ferrari, 2015; Złotowski et al., 2015).

Belief vs. Make-Believe

One profound issue is often overlooked in the discussion of users' expectations of human-like robots and the operationalization of their attributions of animacy to robots: That of whether users' behavioral and linguistic expressions of animacy attributions are founded in an actual belief that the robot in question is animate, maybe even driven by human-like intentions, or whether these expressions are merely metaphorical ascriptions, a performance of "make-believe", of "as-if the robot were alive".

We can find this distinction in several theoretical approaches to the attribution of animacy, agency, and intentionality to technological artifacts (Section 2.2 will discuss these terms in depth). John Searle (1983), for example, distinguished "intrinsic intentionality", which is based on existing mental states of a conscious living being, from ascribed "as-if intentionality", which is used in a metaphorical way to explain the actions of inanimate objects. Similarly, Epley and colleagues (2007) distinguished between "strong" and "weak" anthropomorphism. "Strong anthropomorphism" would entail the explicit belief that a nonhuman entity has humanlike characteristics, for example in the context of religious belief. In contrast, the metaphorical ascription of human likeness to artifacts known to be inanimate would be a form of "weak anthropomorphism". Eleanor Sandry (2015a, p. 11) used the term "tempered anthropomorphism" in a similar vein, meaning the "human understanding … of the robot as somewhat humanlike or animal-like, but … continually tempered by also perceiving the robot as a machine". Other authors propose that anthropomorphism can be understood as a spectrum with different shades or levels (e.g. Persson, Laaksolahti, & Lönnqvist, 2000).

Empirical studies in HRI, HCI (human-computer interaction) and HMI (human-machine interaction) research sometimes make distinctions like these. For example, the widely cited Media Equation study observed that users "mindlessly" attributed social attributes to computers – but also explicitly noted that none of the participants actually said that a computer should be understood in human terms or treated as a person (Reeves & Nass, 1996; also see Nass & Moon, 2000). The authors thus carefully ruled out anthropomorphism as a term to be applied to their observations. In the context of human-robot interaction, Leila Takayama (2012) observed different "levels" of anthropomorphic attributions being applied to the same nonhuman artifact and thus proposed to distinguish observers' "in-the-moment" perspective on robots from a "reflective" perspective. In the actual moment of interaction, a user might be quick to perceive a robot's behavior as agentic or even animate – a "visceral" interpretation, which can differ substantially from a more reflective perspective that would explain the robots behavior with the robot's programming.

Most studies do in fact refrain from operationalizing, or even just addressing, the complex, multifaceted nature of anthropomorphic attributions. This draws criticism from within the HRI community:

> "In the large body of experimental work on human reactions to anthropomorphic robots, responses on standard questionnaires are commonly taken to demonstrate that subjects identify a robot's displays or movements as … expressions of the fundamental human emotions. … Taking these responses … at face value ignores the possibility that they are elliptical for the subjects' actual views. … Saying that the robot has a 'happy' expression might be shorthand for the claim (for example) that if the robot were a human, it would have a happy expression." (Złotowski et al., 2015, p. 348)

The research discussed above shows that "metaphors that might represent a very weak form of anthropomorphism can still have a powerful impact on behavior" (Epley et al., 2007, p. 867), and that the power of "weak" anthropomorphism, of the "merely" linguistic and metaphorical attributions of animacy to technical artifacts, should not be underestimated. In scenarios of human-robot interaction, humans' ability to temporarily suspend their disbelief (Duffy & Zawieska, 2012) or even simply to "perform" the belief of a robot's animacy (McGonigal, 2003; cf. Jacobsson, 2009) can serve as a crucial facilitator for a smooth interaction. Anthropomorphic metaphors can serve as linguistic devices allowing efficient communication about technological artifacts – a "convenient fiction … that permit[s] 'business as usual'" (Caporael, 1986, p. 218). This is especially relevant for complex and difficult to grasp technologies:

> "To confront the relatively unknown in an infinitely complex reality, we must rely upon our understanding of the relatively familiar. The resulting metaphorical concepts help organize inquiry and interpretation – they are necessary [and] fruitful." (Krementsov & Todes, 1991, p. 68)

We are very well able to understand metaphors as what Paul Ricoeur (1978, 2003) called "split reference", interpreting them simultaneously in a literal way, and as an imaginative concept. Nonetheless, it remains difficult to distinguish clearly between the playful, even useful, use of metaphors, the suspension of disbelief as an enabler for smooth human-robot interaction, and potentially harmful misunderstandings about the actual animacy of a technological artifact. After all, "every metaphor is the tip of a submerged model" (Black, 1979; cited in Watt, 1997, p. 60), and talking about a robot as if it were alive might correspond to having a mental model of a robot being a living being.

Are robot designers therefore guilty of deceiving users when they give robots human-like characteristics, when "robots are designed in such a way

that they trigger us to 'fool ourselves'" (Turkle, 2011a, p. 20)? This question has been raised by several actors in the HMI and HRI community (e.g. Borenstein & Arkin, 2019; Coeckelbergh, 2018; De Graaf, 2016; Scheutz, 2012; Sparrow, 2002; Sparrow & Sparrow, 2006). Karolina Zawieska (2015) argues that the core of anthropomorphism is illusion and the topic therefore intrinsically tied to ethical concerns:

> "The main ethical issue lies not in deception itself but rather in a particular view of man where human beings are seen as creatures whose anthropomorphic projections can be evoked 'automatically' and their interaction with robots fully managed and controlled." (Zawieska, 2015, p. 1)

We will also encounter this discussion of deception, in varying forms, at the stops of our empirical tour along the life cycle of robots in the following chapters, and will revisit it once again the final discussion in Chapter 6.

In conclusion, animacy attributions – for example in the form of "anthropomorphic projections" – are a complex and controversially discussed issue in the HRI community. In the context of HRI studies, the focus of academic interest is almost exclusively on the actual or potential interaction between a robot and a human user. However, this moment of interaction is only one very narrow "slice" of the whole life cycle of robots. In the following chapters, we will see that animacy attribution is also an influential phenomenon in all other stages of the cycle. In three exemplary explorations – of robotics engineering practice, of demonstrations, science communication and marketing, and of media discourse on robotics – we will encounter different forms of animacy attribution, and explore its context-specific constructive role.

2.2. Terminology: Anthropomorphism, Agency, Animacy, and More

Before we continue our tour along the life cycle of robots, we first must clarify some of the terminology used in this book. This section will tease apart several overlapping concepts – such as animacy, agency, and intentionality – and it will establish "attribution of animacy" as a central term for this book.

In the vast body of scientific literature on human-robot interaction (cf. Section 2.1) the term used most often for the phenomenon of humans ascribing lifelike qualities to robot technology is "anthropomorphism" – meaning "the attribution of human traits, emotions, or intentions to non-human entities" (OED, n.d.-d). Its derivation from the Greek "ánthrōpos" ("human") and

"morphē" ("form") points to a crucial limitation of the term. By definition, it refers to the attribution of human characteristics to something. In the context of robotics and HRI, however, anthropomorphism is often used to mean something else. Firstly, the term is often (mis)used to describe the human-like design or behavior of a robot, instead of the phenomenon of attribution (Julia Fink, 2014, p. 63; cf. Bartneck & Forlizzi, 2004). This disregards that a robot "is not anthropomorphic per se, but only in so far as it gives rise to anthropomorphic processes in a given user and situation" (Persson et al., 2000, p. 1). Secondly, the term anthropomorphism is frequently used to describe a much wider phenomenon: the attribution of characteristics of living beings in general to robots. Characteristics such as aliveness, emotionality, personality, and sociality are not unique to humans, but apply to a much wider group of living entities. Rarely, the term "zoomorphism" is used for the attribution of characteristics of nonhuman animals to robots. A "zoomorphic robot" is usually understood as a robot with an animal-like design.

In the following chapters, we will encounter several instances of features being attributed to robots that sometimes are characteristic to living beings in general (such as sensory experiences, intentions, or emotions) and sometimes are more specific to human beings (such as long-term life goals). In some existing concepts, this phenomenon is understood to be one level of anthropomorphism (e.g. Persson et al., 2000). But for the purpose of this book and the phenomena it describes the wider term "attribution of animacy" is more adequate.

"Animacy" is a grammatical and semantic feature meaning the "the quality or condition of being alive or animate" (OED, n.d.-a), the adjective "animate" meaning "endowed with life, living, alive" (OED, n.d.-b). Their antonyms "inanimacy" and "inanimate" will also play a role in this book. Animacy also happens to be used in the cognitive sciences and developmental psychology in the context of research exploring, for example, the perceptual and attentional processes involved the identification of living entities in our visual environment (cf. Section 2.3). With research in HRI drawing heavily on the cognitive sciences (cf. Section 2.1) the term animacy also made its way into the robotics literature. However, animacy, with its connection to animism, comes with a difficult colonialist connotation, which is rarely discussed reflexively, or even acknowledged, in the HRI and cognitive science literature (cf. Section 2.3).

Despite this connotation, this book will use the term "attribution of animacy" for the phenomenon in the focus of interest – for two reasons: Firstly,

"animacy" is used in the majority of the relevant HRI literature. Another possibly adequate term – "aliveness" – has only been used by a handful of authors (e.g. Turkle, 2010; Sandry, 2018). Secondly, "attribution of animacy" can be understood as something like the lowest common denominator of the different variations of the phenomenon this book explores. A more confined term like "anthropomorphism" would not adequately reflect the observation that people also ascribe physiological processes – which are not unique to humans – to robots.

The term "attribution of agency", too, would be too restrictive for the context of this book. It is, however, important to acknowledge the importance and relevance of the concept of agency. Depending on the disciplinary context (sociology, philosophy, cognitive sciences …), definitions of agency focus on slightly different aspects. At the most basic level it is "the at least partially independent capacity to engage in goal-directed action" (H. M. Gray et al., 2007).

At this point, it is important to note that

> "the concepts of 'animacy' and 'agency' … are not coextensive. Animate entities are living things that can act as agents …. Living things that are not sentient and do not act as agents, such as trees and mushrooms, are not animate. The domain of agents, however, can include inanimate automatons, such as robots, that generate their movements and actions to achieve goals." (Gobbini et al., 2011, p. 1911)

Science and technology scholars have long been discussing whether non-biological entities can possess agency (also see Section 2.3). Werner Rammert (2008) proposed a multi-level model of agency. On the model's lowest level (causality), agency means simply "behavior that exerts influence or has effects". This level, on which "it doesn't make any difference whether humans, machines or programs execute the action" (Rammert, 2008, p. 11), has obvious parallels to the concept of generalized symmetry within Actor Network Theory: "Objects too have agency" (Latour, 2005, p. 63). The next higher level (contingency) requires the capacity to act differently, to choose among several behavioral options. Only the third level uses the term intentionality, referring to reflexive and intentional actions. Rammert (2008, p. 12) argued that technical artifacts, while not able to have "literal" intentionality, "can be constructed as if they had an intentional structure". Chapter 4 (Section 4.7) will explore in more depth the issue of technical artifacts acting "as if" they were animate.

In the cognitive sciences and HMI literature one can find many more proposals for the conceptual relations of the terms discussed in this section. For example, Heather Gray and colleagues (2007) understand agency – defined here as the capability to act and intend – as one dimension of "mind" that can be attributed to an agent or entity (next to the dimension of experience, i.e. the capability for feelings and sensations). Elsewhere, Florent Levillain and Elisabetta Zibetti (2017, p. 13) propose that the behavioral cues of autonomously acting technological artifacts are interpreted by an observer on three levels: the Animacy Level (Does the object look alive?), the Agency Level (Does the object appear to act intentionally?), and the Mental Agency Level (Does the object appear to take into account others' goals?).

The terms discussed above (anthropomorphism, animacy, animism, agency, intentionality…) are those one encounters most frequently in the current academic literature. There are also less frequently used concepts, such as "Universal Projection" – used by Thomas Luckmann (1983) to describe humans' capacity to project their own living body onto everything they encounter in the world, which is sometimes referenced in the context of HRI (e.g. Nørskov, 2017, p. 11). As is "Mythopoeic Thought", a proposed ancient form of human thought, in which each observed event is attributed to the will of a personal being (Frankfort et al., 1946; referenced e.g. by J. Carpenter, 2016, p. 20).

2.3. Disciplinary Perspectives: Animacy Attribution as an Object of Research vs. Methodological Malpractice

While HRI is the academic field where most research on animacy attributions takes place at the moment, the issue is also of interest for many other disciplines. An exploration of the publications of different academic fields reveals two overarching perspectives: Firstly, animacy attributions as a methodological malpractice, and secondly, as an object of research in itself.

For centuries, scientists freely compared natural phenomena to processes of the human body and mind. Medieval scholars attributed chastity to camels and self-sacrifice to storks, renaissance scholars referred to nature as a benevolent servant or artist (Daston, 2000, p. 29). By the seventeenth century, however, natural philosophers started to abandon these comparisons. The explaining of natural processes with human-like beliefs and desires became to be considered scientific misconduct: "Nature had become irretrievably 'the

other'" (Daston, 1995, p. 38, cf. 2000). It took a while for this perspective to reach the non-academic community. Up until the nineteenth century, fed by the reports of travelers, naturalists, and amateur scientists, zoopsychological publications describing animal behavior with "human" terms stayed wildly popular, "replete with descriptions of 'states' and 'factories', 'art' and 'crafts', … 'friendship', 'wars' … among animals" (Krementsov & Todes, 1991, p. 76). By the end of the nineteenth century, criticism of the zoopsychological perspective on animals reemerged and most scholars agreed that "in no case may we interpret an action as the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale" (Morgan, 1894, p. 53). In the 1960s, researchers studying great apes – among them Jane Goodall – were strongly criticized for attributing presumably human characteristics, such as emotions, to animals – the "worst of ethological sins" (Goodall, 1993, p. 15; cf. Rees, 2001).

Animism – the attribution of life or spirit to nonliving entities (OED, n.d.-c) – was traditionally viewed as an immature disposition. As such, it stayed of interest mainly in two academic contexts: On the one hand, research in developmental psychology on certain phases of infants' cognitive development (e.g. Piaget, 1929). On the other hand, early anthropological research on "primitive" religions ascribing a distinct spiritual essence to objects, places, and creatures (cf. Franke, 2010). This "old animism" perspective of anthropology, viewing "primitive" animist cultures as being unable to differentiate between persons and things, was held, for example, by nineteenth-century anthropologist Edward Burnett Tylor (e.g. 1871). Today it is criticized for its colonialist and dualist worldviews and rhetoric (Harvey, 2006, p. xii). Nonetheless: "images of fetishes, totems, … tribal art, pre-modern rituals, and savagery … have forever left their imprint on the term [animism]" (Franke, 2010, p. 11). Up until recently, practices of attribution of animacy to non-living entities were regarded in most scientific disciplines as both an archaic or infantile reflex and as a methodological mistake (Vidal, 2007, p. 3). As something that scientists knew having to avoid at all costs for its "violat[ion] of the ideal of the objectivity of perspective"[16] (Daston, 2000, p. 28), little scientific attention was directed to the nature and consequences of animacy attributions for a long time:

16 Translated from German by the author.

> "The debate about the nature and implications of anthropomorphism has rarely been neutral or scientifically objective but has focused mainly on its fallacious essence … which has diverted attention away from the goal of understanding the nature of the phenomenon." (Urquiza-Haas & Kotrschal, 2015, p. 168)

Only in the last few decades, fueled by observations of human interactions with increasingly complex and autonomous technologies, scientific interest reemerged across academic disciplines (cf. Vidal, 2007). For example, communication scientist Sherry Turkle, based on her ethnographic research on computer users, proposed that computers were more than "just a tool" and explored how we interact socially with them (Turkle, 2005, p. 3). She would later coin the term "evocative objects", describing how certain machines "can act as a projection of part of the self, a mirror of the mind" (ibid., p. 20) and can even become emotional and intellectual "life companions" (Turkle, 2011b, p. 9). Similarly, communication scientists Byron Reeves and Clifford Nass showed with their Computers as Social Actors paradigm that even minimal social cues from a technical artifact can cause humans to mindlessly treat it like a living interaction partner (e.g. Nass et al., 1993). Their observations are often referred to as the Media Equation, after the title of their widely cited book (Reeves & Nass, 1996). In a series of HCI studies they showed that human "individuals are responding mindlessly to computers to the extent that they apply social scripts – scripts for human-human interaction – that are inappropriate for human computer interaction" (Nass & Moon, 2000, p. 83). Communication scientist Don Ihde (1990, p. 97 ff.) explored the interaction of humans and machines as a quasi-other, proposing "alterity relations" as a term for relations with technology (cf. Sandry, 2018).

In the cognition science community, a widely cited study from 1944 by Fritz Heider and Marianne Simmel showed that human subjects interpreted movements of abstract shapes in an animation film as social interactions between animate entities. For decades, this study was mainly perceived as an interesting anecdote (Aarts, Dijksterhuis, & Dik, 2013). In recent years, however, it has been replicated several times and is now regarded as seminal for the research of social perception and causal attribution (cf. Lück, 2006). Today, there is a lively research community interested in the cognitive processes and neural structures involved in the perception of animacy and action – both in the developing and adult brain (e.g. Gobbini et al., 2011; Marsh et al., 2010). The ability to identify animate entities is already present in infants

(Kuhlmeier, Wynn, & Bloom, 2003; Woodward, 1999). It is understood to be the foundation for the later development of a Theory of Mind – the ability to attribute internal mental states to others (Premack & Woodruff, 1978). Research in the cognitive sciences found that agentic entities in our visual field are prioritized via attentional selection, compared to inanimate objects (e.g. New, Cosmides, & Tooby, 2007; Scholl & Gao, 2013). Which entity is categorized as animate depends not only on the visual appearance (e.g. the presence of eyes), but also on its behavior. For example, the perception of movement being goal-directed and self-propelled strongly contributes to an entity being categorized as behaving intentionally and having human-like mental states, and to the observer behaving towards the entity as if it was alive (see e.g. Epley & Waytz, 2010, for an overview). John Harris and Ehud Sharlin (2011) explored human reactions to abstract motion with the help of an extremely minimalistic robot consisting of nothing but a stick, which was remote-controlled to perform different movements. Observers not only consistently rated certain movements as emotional expressions (e.g. speed – excitement, approach – aggression), but also spontaneously tried to find meaning in the movements and attributed mental processes to the robot.

A proposed explanation for these reactions is that the cognitive-perceptual subsystem responsible for the identification of agentic entities in our environment is so sensitive that it is prone to over-interpret even minimal perceptual cues. The evolutionary reasoning is that erring in favor of interpreting an object in our environment as animate increases the probability for survival. From an evolutionary perspective, being able to detect other intentionally acting agents in our vicinity is a crucial fitness advantage. Being able to quickly identify a predator can mean the difference between life and death (e.g. B. J. Ellis & Bjorklund, 2005). Also beyond the immediate threat of being killed by a wild animal, humans, as highly social animals, have been profiting from this ability in the context of their complex social lives – for example when establishing alliances with other human tribes. This idea was conceptualized in the so-called Social Intelligence Hypothesis (Kummer et al., 1997). The idea of a "Hyperactive Agency Detection Device" is even proposed as an explanation for religious beliefs in a higher power (Barrett & Lanman, 2008):

> "Based on stimuli in the moment, we ascribe the highest level of sophistication possible to the object at hand. … The smallest evidence of live or intentional action encourages perceptional shift, allowing us to ascribe live and intentional statuses to objects more readily." (Owens, 2007, p. 573)

Animacy as a default interpretation of ambiguous stimuli has been proposed by several researchers. For example, Daniel Dennet (1998) postulated that the "Intentional Stance" is the most abstract of three possible levels of abstraction[17] when considering the mental state of an entity. When taking the Intentional Stance, predictions made for the behavior of an entity are based on its assumed beliefs and desires – compared to, for example its physical properties, respectively its design purpose, on the two less abstract levels. Similarly, Stewart Guthrie (1997) proposed an "involuntary perceptual strategy", and Linnda Caporael and Cecilia Heyes (1997) a "cognitive default", in that "we will default to human characteristics whenever going gets rough" (ibid., p. 64). Within the cognitive sciences, the phenomenon of animacy attribution is considered "endemic" (Watt, 1997, p. 125), "almost irresistible" (Eddy, Gallup, & Povinelli, 1993, p. 88), and "inevitable" (Krementsov & Todes, 1991, p. 80), and researchers are trying to "set traps" (Caporael, 1986, p. 217) for it, in order to "tame" it for research and application development – as discussed for the context of HRI studies in Section 2.1 of the present chapter.

In the 1980s and 1990s, the field of science and technology studies (STS) began to explore the agentic and interactive role of technological artifacts. It was the context of scientific practice where STS researchers first began to explore the crucial impact of non-human artifacts – such as microbial samples or scientific instruments – on practices of scientific knowledge production (e.g. Knorr-Cetina, 1981; Latour & Woolgar, 1986; Lynch, 1985). This research resulted in a re-conceptualization of the prevailing ontological separation of "the social" and "the technical" into a concept of human and technical agency existing in parallel (e.g. Bijker & Law, 1992; Latour, 2005; Law, 1991; MacKenzie & Wajcman, 1999; also see Krummheuer, 2015). In a "turn to technology", researchers began exploring the social shaping and construction of technology (Woolgar, 1991). Rather than looking at the impact of technology on society, this research was – and still is – interested in how societal context finds expression in technological developments, exploring ideas such as material agency (cf. Knappett & Malafouris, 2008) and artificial interaction (e.g. Braun-Thürmann, 2002). Lucy Suchman (1987, 2007) described new human-machine configurations and human interaction with intelligent machines in her "Plans and Situated Actions". Karin Knorr-Cetina (1997, 1998) proposed the concept of a "sociality with objects" after observing human-object relationships with a perceived mutuality and solidarity.

17 Physical Stance, Design Stance, Intentional Stance.

Crucially, while this research on "autonomous technology [took] place down on earth … it also influence[d] the higher spheres of philosophical debates about the ideas of agency and autonomy" (Rammert, 2011, p. 1). Actor Network Theory (ANT) proposed a radical symmetry of human and nonhuman actors ("actants"), meaning that both are fundamentally equal in their contribution to any effects they have on the environment (cf. Section 2.2). In coming together in heterogeneous networks, human and nonhuman actants are presumed to constitute sociotechnical ensembles, which, as a whole, serve as the location of any agency and create meaning in the world (e.g. Latour, 1987, 2005; Callon, 1986). In contrast to ANT's approach, Werner Rammert and Ingo Schultz-Schäffer (cf. Section 2.2) suggested a distribution of agency between humans and technical artifacts, with the attribution of agency to human or nonhuman agents being constructed only by the observer (Rammert, 2002, 2008, 2011; Rammert & Schulz-Schaeffer, 2002a).

Already in the next chapter, we will encounter such a perceived distribution of agency – between roboticists and "their" robots. Also in the following chapters, while exploring a range of practical and discursive human-robot interaction, we will revisit and apply many of the conceptual approaches discussed above.