A ladder of agency in perceived animacy

Tao Gao[1], Ning Tang[1]

[1] University of California, Los Angeles

Abstract

This chapter investigates how perceived animacy fundamentally drives our understanding of complex social interactions, exemplified by the classic Heider & Simmel displays. Empirical findings from perceptual, developmental, and linguistic studies are structured into a "ladder of animacy," outlining increasing levels of agency that reflect progressively richer mentalization: (1) rapid motion, (2) self-propelled motion, (3) utility-maximizing motion, and (4) motion driven by intentional commitment. Evidence across these fields suggests animacy as a core concept that shapes how we spontaneously perceive the world, develops early in infancy, and serves as the non-linguistic basis for verb semantics. However, the chapter also identifies research gaps, such as the adaptability of the rationality principle in perception and conceptual differences in causal-physical interactions across perception and linguistics. The chapter concludes by underscoring the importance of understanding agents' intentional actions in alignment with verb semantics, bridging insights from perception, development, and linguistics.

Keywords: perceived animacy, agency, Heider & Simmel displays, cognitive semantics, core knowledge, utility maximization, intention

1 Introduction

“I hope Artificial Intelligence (AI) can make my life easier by doing laundry, chores, washing dishes and walking the dogs, but instead it threatens my job by writing poems, essays, and codes”.

This lament vividly revealed the unbalanced development of AI, compared to human intelligence: On the one hand, AI, as represented by the large language models trained on huge corpuses of text data. On the other hand, AI achievements on non-linguistic intelligence, with the causal, physical, and social commonsense knowledge to interact with the objects and social agents in the real world—rather than the language description of it—remain much more limited. It was first pointed out by Hans Moravec that “it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility” (Moravec, 1988).

With the rapid advancement of AI and the continuous breaking of technological barriers, we are closely watching whether the Moravec paradox will remain relevant in the near future. Currently, this paradox is well illustrated by a classic phenomenon: the Heider-Simmel displays (Heider & Simmel, 1944), where humans irresistibly interpret the motion of simple geometric shapes as animate agents with desires and intentions, mapping their actions onto a rich set of verbs such as ‘chase’, ‘follow’, ‘fight’, and ‘hesitate’. Despite significant progress in computational modeling of the Heider-Simmel displays using advanced AI technologies such as Bayesian hierarchical models (Shu et al., 2021) and graph neural networks (Malik & Isik, 2022), no model can yet

replicate what humans do intuitively—transforming impoverished visual stimuli into rich, story-like linguistic descriptions of social activities.

Studies of the Heider & Simmel displays have interested researchers from many subfields of cognitive science, including perception, infant cognition, and psycholinguistics. To organize this vast amount of literature with diverse perspectives and methods, we adopt two theoretical frameworks: one that outlines the cognitive architecture supporting perceived animacy, and the other that focuses specifically on the mechanisms of animacy.

For the cognitive architecture that can host perceived animacy, we adopted the conceptual space hypothesis proposed in the field of cognitive semantics, which emphasizes synergies between perception, infant cognition, and linguistic semantics (Jackendoff, 1985). First, the contents of the conceptual space are derived from the outputs of perception, which is not a replica of the real world but actively constructs continuous, analog sensory inputs into discrete, symbolic mental entities. Humans can consciously experience these projected mental entities as the output of perception in the conceptual space; however, they have no conscious access to how the projection is achieved within perception. The ontology of these projected entities includes objects, causality, physics, paths, events, and agency.

Second, one may notice that the ontology of the conceptual space highly overlaps with the domains of core knowledge from infant studies (Spelke & Kinzler, 2007), naturally leading to the assumption that basic ontological categories in the conceptual space develop early in infancy. Third, it is noteworthy that the conceptual space was proposed primarily as a theory of semantics

in linguistics, not as a theory of perception and infant cognition. The key theoretical insight is that there is no separate semantic space—the conceptual space is the semantic space. Linguistic semantics can be broken down into non-linguistic concepts that shape syntax through linking rules. While the linguistic theories of these linking rules were well developed at the time the conceptual space was proposed, the evidence supporting its non-linguistic nature was largely speculative. In fact, in discussing how sensory inputs are projected into the conceptual space, the primary evidence was the classic Gestalt principles. Yet, in the last decades, this theory has been greatly supported by a wide range of perception studies inspired by theories of semantics, such as noun versus substance distinctions in object recognition (Scholl & Pylyshyn, 1999; Soja et al., 1991; VanMarle & Wynn, 2011), types of causality (Leslie & Keeble, 1987; Michotte, 1963; Talmy, 1988; Wolff, 2003), and more recently, event boundaries (Baldwin et al., 2001; Papafragou & Ji, 2023; Strickland et al., 2015). Therefore, we apply this theoretical framework to studies on animacy to review evidence from the above three fields, aiming to identify the synergies (or the lack thereof): that the understanding of animacy develops early in infancy, is automatically included as part of human adults' visual experience, and is reflected in verb semantics and syntax.

When discussing the specific mechanisms of perceived animacy, we face the challenge of organizing a wide range of literature with diverse theoretical and methodological focuses. Some studies concentrate on identifying salient behavioral patterns of animate agents and the visual cues corresponding to these properties (Barrett et al., 2005; Dittrich & Lea, 1994). Others explore concise principles of agency that can be applied across various actions and environments (Gergely et al., 1995), examining the mental representation of agency with different levels of

complexity (Cheng et al., 2023). In this chapter, we address this challenge by addressing the fundamental question: What is an animate agent? The answer to this seemingly simple question turns out to be surprisingly complex. Instead of pitting these perspectives against each other, we propose a ladder of agency as a framework to discuss the levels of agency addressed by different kinds of evidence. Before we delve into this ladder, we would first like to clarify what our framework is not.

First, we do not adopt a distinction between animacy and agency, and we will use these two terms interchangeably. A common view is that animacy could be more primitive, concerning the dichotomy between inanimate object and animate object. Once an object is categorized as animate, then a subsequent processing of agency can be triggered, to further explain the motion of an animated object with more sophisticated mental states. At least in the field of visual perception, we have not seen compelling evidence supporting this two-stage sequential processing. Therefore, we do not adopt a hard distinction between these two related concepts and assume perceiving animacy is perceiving agency—how the observed motion is driven by the agent's internal mental states.

Second, the idea of a ladder of perceived animacy is inspired by scientific studies of agency, in which theorists have proposed various ladders of agency that rank the levels of intelligence an agent possesses. For example, Dennett proposed a hierarchy comprising Darwinian creatures, which rely on instincts endowed by evolution; Skinnerian creatures, which learn from trial and error; Popperian creatures, capable of predicting the future by running internal simulations; and Gregorian creatures, empowered by mental tools inherited from culture (Dennett, 1996). In

parallel, Pearl introduced a ladder with three levels: Observing, where agents learn from statistical associations; Doing, where agents change observed statistical patterns by intervening through their own actions; and Imagining, where agents engage in counterfactual reasoning about what could have happened in alternative worlds (Pearl, 2009). While these are deeply insightful scientific theories of agency, they do not offer a framework for understanding perceived animacy. It is unlikely that when viewing the Heider & Simmel displays, people are concerned with distinctions between Skinnerian and Popperian creatures. Therefore, we propose a different ladder of agency based on the complexity of mental states attributed to the agent when viewing its motions. Unless otherwise specified, when we refer to a theory of animate agents, we mean the folk theory of animacy adopted by the human intuitive conceptual system, not the scientific definition of a biological agent. That being said, if our perceptual system has evolved to efficiently represent other agents' mental states, the theory of agency underlying our perception may align with certain aspects of scientific models of agency, as we will demonstrate later.

2 A ladder of perceived animacy

2.1 Limitations in the commonsense definition of animacy

Before we articulate the ladder of perceived animacy, we want to first point out why the dictionary definition of animacy is a poor guide for reviewing research on the Heider & Simmel-like displays. Animacy is typically defined as the quality of being alive (Oxford English Dictionary) or sentient (Wikipedia contributors, n.d.). This definition is simultaneously being too broad and too constrained. It is too broad because “be alive” means that a plant should be

categorized as animate, which is not consistent with human intuition, as summarized in Westfall (2023). For example, young children can identify unfamiliar animals as animate but often do not group plants with animals. In addition, they recognize animacy in animals before they understand that plants are also living creatures (Opfer & Gelman, 2011). Adults, including experts like biology professors, often do not spontaneously categorize plants as alive under time pressure (Goldberg & Thompson-Schill, 2009). Neural studies also indicate that plants are processed more similarly to artifacts than to animals in the brain, as indicated by object representation dissimilarity in the inferior temporal cortex, where the response pattern for plants and artifacts is more similar than that for plants and animals (Kriegeskorte et al., 2008). In the active field of vision research on perceived animacy, no one is even bothered to discuss why plants do not count. As we are guided by the synergy between vision and language through the hypothesized conceptual space, we want to point out that the exclusion of plants from animacy is also supported by cross-linguistic studies. Across diverse languages, there is a common tendency to assign higher animacy status to animals than to plants within grammatical systems (Yamamoto, 2006). This hierarchy influences various aspects of grammar, such as noun classification, verb agreement, and object marking. These cross-linguistic tendencies suggest a universal cognitive bias in how humans perceive and linguistically encode living entities, positioning animals above plants on the animacy scale. This reflects a shared aspect of human cognition in distinguishing between different forms of life. However, it's important to note that exceptions exist, and not all languages grammaticalize this distinction explicitly.

Conversely, requiring sentience as a criterion for animacy sets the bar too high, as it involves qualities like feeling, self-awareness, or consciousness. For example, I consider a mosquito

flying around me to be animate, yet I have no idea whether it possesses feelings or is aware of its own existence. Moreover, this definition would exclude a wide range of highly capable AI agents, such as robotic vacuum cleaners or autonomous cars, whose intelligent behaviors certainly merit our attention, even though self-awareness or consciousness are irrelevant to the tasks they perform. Russell (2019) highlights the fallacy of assuming sentience as a prerequisite for intelligent agency. He observes that in many science fiction movies featuring malevolent AI intent on destroying humanity, the narrative often begins with the AI becoming sentient. In reality, safety concerns about AI have little to do with it achieving consciousness; rather, AI capabilities are rapidly advancing through improvements on specific task benchmarks, such as computer vision models in the accuracy of object recognition and large language models in test scores for taking various exams, not through a pursuit of sentience.

Therefore, we define animacy as the capacity to observe the environment and to act on it rapidly based on its own mental states. This definition offers two key advantages. First, it avoids ambiguous terms like 'alive' or 'sentient,' whose definitions are inherently vague and difficult to pin down precisely. Second, it allows a ladder of agency, where the sophistication of observation-action mechanisms determines an entity's position on this ladder.

2.2 Ladder 1: Rapid motion

Why are plants not typically conceptualized as animate, despite being living organisms that respond to their environment? Plants exhibit behaviors such as roots growing toward water and

sunflowers orienting themselves to the sun, which show responsiveness. Moreover, people can easily attribute intentionality to plant behaviors, saying things like "the tree is *trying* to penetrate deep into the soil" or "the flower is *catching* the sun" (Dennett, 1989). We even describe seeds as "being cautious" about when to sprout in spring. Yet, despite these rich interactions with their environment and our capacity to mentalize their behaviors, plants are not perceived as animate beings in the same way animals are.

The key distinction lies in temporal scales of movement: animals respond within seconds, as is the case in most human psychophysical experiments, while plants operate on timescales of hours. Such slow movements fall outside the range of human online visual perception. This perceptual constraint shapes our conceptual space, creating fundamentally different representations for plants and animals. These distinct representations, in turn, influence both cognitive development and semantic processing. This interpretation is supported by viewer experiences with scientific documentaries such as Planet Earth, which frequently accelerate plant footage to show processes like flower blossoming in seconds rather than hours. At these increased speeds, not only can viewers grasp hours-long processes in seconds, but the plants also appear more animate.

The need for speed in recognizing animacy suggests that the distinction between animacy and non-animacy lies in whether an object can act in real time, rather than whether it is alive. This is likely because of the evolutionary pressure for us to interact with animals around us in real time. Perception, with its focus on fast processing of continuous visual information, is well-suited for this task. But what is real time? Studies on time perception have revealed that our perception of

"now" is not a fleeting, infinitesimal moment but rather spans a small window of approximately three seconds, known as the specious present (see Pinker, 1997; Pöppel, 1997, for a review). This interval captures not just the immediate moment but also a bit of the recent past and the near future. It is significant because it aligns with the time required for intentional actions like waving to someone across the street or the rapid planning of precise movements such as pouring a cup of coffee without spilling. Importantly, it is also the duration during which information can be reliably stored in working memory without significant decay. In other words, our experience of the present and the robustness of our working memory are perfectly suited to processing intentional actions—either generating them ourselves or perceiving them in another agent.

The engagement of working memory in maintaining a brief period of rapid motion for perceiving animacy has been supported by recent vision science studies. One study (Wick et al., 2019) using Heider-Simmel style animations found that observers' ability to match narratives to scenes declined sharply when the number of moving items exceeded three, with the most dramatic drop occurring between three and four items–roughly identical to the capacity of visual working memory for memorizing (e.g., Luck & Vogel, 1997) or tracking (e.g., Scholl & Pylyshyn, 1999) objects. Another study directly tested how motion information accumulates in working memory during the perception of chasing by combining human psychophysics and Bayesian cognitive modeling (Gao et al., 2019). Results indicated that human performance in detecting chasing among randomly moving distractors aligns best with a capacity-limited Bayesian model that tracks up to four objects and accumulates motion information over 2-3 seconds in working memory. Allowing motion trajectory to decay too quickly or slowly causes the model to significantly under- or over perform relative to human results. These findings suggest that rapid

motion is essential for perceiving animacy: when motion is too slow, information critical for inferring mental states exceeds the three-second working memory window, preventing adequate accumulation of data needed to identify intentional actions.

2.3 Ladder 2: Self-propelled motion

Clearly, rapid motion is necessary but not sufficient for perceiving animacy. A theory of perceived animacy must specify which types of rapid motion are indicative of animate beings. We begin with the simplest model: an object that makes abrupt and unpredictable changes in its direction of motion is perceived as an animate agent. This perception arises because an inert, inanimate object's motion—or lack thereof—can typically be fully explained or predicted by the physical environment. Examples include changes in motion direction upon hitting another object or the predictable path of free fall when unsupported. Therefore, if an object exhibits an unpredictable change not explainable by external physical forces, it must possess an internal energy source, making it an agent capable of autonomous action (Gelman et al., 1995).

The self-propulsion theory of animacy is supported by infant studies. The sensitivity to the distinction between inert and self-propelled objects may exist in newborns, as evidenced by their visual preference for motion that originates from rest through self-propulsion, rather than for motion that is already in progress upon appearing on the screen (Di Giorgio et al., 2017). Additionally, when presented with objects moving at a constant speed versus those exhibiting sudden speed variations, newborns showed a visual preference for objects that accelerated and

then decelerated but displayed no such preference for objects that only accelerated or only decelerated (Di Giorgio et al., 2021). By 5 to 6 months of age, infants have an intuitive understanding of physics, recognizing that inanimate objects cannot move on their own without physical contact. (Leslie & Keeble, 1987). Furthermore, they develop different expectations for inert versus self-propelled objects (e.g., Luo & Baillargeon, 2005). In one study, infants watched an object moving in and out of an occluder (Luo et al., 2009). In the inert condition, the object's motion was launched by a human hand. In the self-propelled condition, the object initiated its motion without any contact. The occluder was then lifted showing a wall, whose position could or could not explain the object's turning back by a physical collision with the wall. Infants displayed surprise only when the object in the inert condition could not have realistically contacted the wall due to its position, suggesting an unexpected change in motion without collision. In contrast, they exhibited no such surprise with the self-propelled object, underlining their understanding that self-propelled entities can reverse direction on their own.

At 7 months, infants exhibit the ability to form category-specific associations between static and dynamic attributes for animates and inanimates (Träuble et al., 2014). They match motion patterns that involve changes in speed and direction in an animate manner with an animate identity of the moving shape, while associating motion along a straight path at a constant speed (i.e., in an inanimate manner) with an inanimate identity. By 9 months, infants develop a more refined understanding of self-propulsion, being able to form different expectations for robots and humans (Poulin-Dubois et al., 1996). They consider self-propulsion by a small robot to be anomalous, whereas self-propulsion by a human stranger is not perceived as unusual.

Vision studies have investigated self-propelled motion as a cue for animacy in human adults. One study found that participants rated a single object as more animate when its motion involved

significant changes in speed and direction (Tremoulet & Feldman, 2000). However, the rating of animacy dropped when the changes in direction could be explained by interactions with the environment, such as collisions with obstacles like a "paddle" (Tremoulet & Feldman, 2006). Another study, using visual tracking paradigm, explored how self-propelled motion affects attention. In this experiment, animate and inanimate motions were defined by whether an object's abrupt change in motion direction could be accounted for by physical collisions. After some time, one of the original objects—either an animate or inanimate object—would disappear. The reaction time measurement showed that participants detected the disappearance more quickly when the disappearing object had previously exhibited animate motion, suggesting that animate motion has a higher priority in visual perception and more effectively captures attention (Pratt et al., 2010).

The self-propelled model captures a straightforward fact: all animate agents have their own energy source, while inanimate objects don't. The motion cue for identifying such energy sources is simple: unpredictable changes. Yet, its theoretical implications are somewhat confusing, perhaps because as a model of agency, it is too simple. The model occupies an ambiguous position at the intersection of physics and animacy, making it unclear which domain it truly belongs to. A clear sign of this confusion is the difficulty in naming the theory. Researchers often describe self-propelled motion as a "violation of Newtonian physics," but they quickly clarify that, in reality, nothing actually violates Newtonian physics—even self-propelled motion can be fully explained within its framework (Tremoulet & Feldman, 2000). The exact amount of force exerted by a hidden energy source can be precisely determined using Newtonian equations of motion (Todorov et al., 2012). In other words, a model that focuses on detecting hidden

energy sources and calculating the forces they exert remains firmly within the realm of Newtonian physics. This raises the question of whether the self-propelled model adequately captures the essence of perceived animacy or if it merely extends physical principles without truly offering insight into how animate agency is represented in perception.

Our position on the confusion surrounding self-propelled motion has two components: we support placing self-propelled motion within the domain of animacy rather than physics; however, it is highly limited as a standalone model of animacy. We fully understand the temptation to describe self-propelled motion as a "violation of physics"—we share this inclination. This temptation reflects a conflict between researchers' intuitive understanding of physics and the scientific knowledge of Newtonian physics. Recent studies have shown that human intuitive physics can be remarkably well modeled as an approximate Newtonian physics engine (Battaglia et al., 2013; Kubricht et al., 2017; Ullman et al., 2017). However, these studies typically involve the motion of inert objects without hidden energy sources, such as judging the stability of a tower stacked with rigid blocks. It is possible that human intuitive physics is constrained to inert objects and excludes the existence of hidden energy sources within them. This assumption becomes evident when one looks at studies on physical knowledge that do not consider animacy. In one study, researchers treat scenarios like a car hovering in mid-air as "violations of physics" (Stahl & Feigenson, 2015), without bothering to clarify that hovering does not violate Newtonian physics if a hidden energy source is assumed. This suggests that, by default, when referring to physics, humans intuitively refer to inert physics. Understanding this helps us appreciate that self-propelled motion occupies an awkward position: it lies outside the realm of human intuitive physics but can be fully explained by Newtonian physics. From this

perspective, it is appropriate to call self-propelled motion a "violation of physics," but it's important to point out that it is human intuitive physics that is violated, not Newtonian physics. The ongoing investigation of human intuitive physics may impact our understanding of perceived animacy, as, after all, a model of agency concerns motion that cannot be fully explained by physics. Recent studies have begun to explore how animacy is perceived when an agent is placed in an environment constrained by physical laws, which has increasingly become a focal point of research (Shu et al., 2021; Tang et al., 2021).

Recognizing that self-propelled motion occupies an ambiguous position between physics and agency also reveals its limitations as a standalone theory for perceiving animacy. One could argue that it is intuitive physics, rather than animacy, doing the heavy lifting in processing self-propelled motion —it grabs attention because it violates predictions of intuitive physics, not because it inherently confirms a theory of animacy. Empirically, it is also questionable how convincingly self-propelled motion conveys animacy. On one hand, the rich social information in the Heider & Simmel displays is universally compelling and hard to ignore. On the other hand, in studies of multiple object tracking (MOT, Scholl & Pylyshyn, 1999), where participants follow several objects moving unpredictably—fitting the definition of self-propelled motion—these objects are studied as examples of object indexing, with little reference to animacy or agency. The reason is clear: despite their random movements, they don’t evoke strong perceived animacy.

Furthermore, self-propelled motion ultimately comes down to motions that are unpredictable. However, this inconsistency becomes evident when we consider the viewing experience of the Heider & Simmel displays, where the agents behave in a highly predictable manner. For most of the video, the agents move in ways that coherently and convincingly reflect social inferences about their mental states. For example, a "victim" predictably hides in a corner, while the "bully" becomes frustrated when tricked into losing its prey. These movements are deliberately structured to align with the inferred intentions and emotions of the agents. Therefore, we argue that self-propelled motion serves as a lower bound or minimal criterion for perceiving animacy. It indicates that animate agents should not behave like inert objects and cannot be fully predicted by intuitive physics. However, this is insufficient as a standalone theory of animacy. A more comprehensive theory should do the opposite—it should precisely predict how an agent will move based on their inferred mental states.

2.4 Ladder 3: Rational action as utility maximization

While a physical model can accurately describe an object's speed, acceleration, and even reveal the presence of a hidden energy source, it cannot explain the underlying reasons—'why' an agent decides to channel that energy into specific actions, exerting force in certain directions and for particular durations. A theory of perceived animacy should focus on interpreting observed motion as an action controlled by an agent’s mental states, which lies clearly outside the realm of physics.

Surprisingly, the theory of agency underlying perceived animacy is well aligned with the standard model of agency in AI (Russell & Norvig, 2016). According to this model, an agent must: 1. Observe the environment so that it can take actions in response to those observations. 2. Have a utility function that assigns a numerical value (utility) to each possible outcome of an action, representing the desirability or preference of that outcome from the agent's perspective. 3. After observing the environment, take actions whose outcome can increase its utility, with a purely rational agent aiming to maximize its utility—this is referred to as the rationality principle. When outcomes are uncertain, the agent chooses actions based on the expected utility—the average of all possible outcomes—which is why the rationality principle is also known as the maximizing expected utility (MEU) principle. Under this definition, a robot vacuum cleaner like a Roomba perfectly fits the description of an agent. It can sense its environment (e.g., detecting debris on the ground), has a well-defined utility function (the cleaner the floor, the higher the utility), and can take actions to increase its utility (it sucks!). In this sense, a Roomba is as much an agent as a dog fetching a ball.

A seminal study demonstrated that 12-month-old infants predict an agent's movements based on the rationality principle, expecting the agent to take the shortest path to a goal (Gergely et al., 1995). After seeing an agent follow the most efficient path, infants anticipate the agent will adapt its actions to a new environment by taking the newly optimal path, rather than repeating an old one that no longer makes sense. For instance, if an agent jumps over a barrier to reach an object, infants expect the agent to move directly toward the object once the barrier is removed, and are surprised if the agent continues to jump unnecessarily. Interestingly, the rationality principle can be applied in the opposite direction: inferring hidden structures of the environment to justify an

observed action (Csibra et al., 2003). For example, if infants see an agent jumping while passing through an occluder, they expect a hidden barrier behind it that would rationalize the jump. Moreover, infants use the rationality principle flexibly, taking into account various contextual cues and physical constraints. When object A moves toward object B and encounters a barrier with a gap, infants interpret A as chasing B only if A is larger than the gap and must detour around the barrier because it cannot pass through. However, if A is smaller than the gap but still detours around the barrier instead of passing through it, infants do not interpret this behavior as chasing B. This demonstrates that the rationality principle can be applied flexibly to interpret a wide range of goal-directed motions, depending on the specific layouts and physical constraints of the environment.

The computational cognitive model of the rationality principle is framed as a Bayesian inference process that identifies the utility function most likely responsible for generating the observed motions (Baker et al., 2009). At its core is a planning engine that calculates the probability of an agent taking a particular action given its utility function p(action|utility), assuming the agent acts to maximize its utility. The planning process is then inverted through Bayesian inference, using the planning engine as the likelihood function (Eq. 1). Combining the action likelihood with a prior over utility functions (which can be a uniform distribution if no prior knowledge is available) allows the model to compute the posterior distribution of the utility function given the observed actions, p(utility|action). Since this Bayesian process effectively reverses the planning engine—taking observed actions as input and outputting a utility function—it is referred to as the inverse planning model.

$$p(utility|action) \propto p(action|utility)p(utility) \quad (1)$$

It is noteworthy that the above equation is the straightforward application of the Bayes rule, just like in the typical data analysis which inverts p(data|hypothesis) to p(hypothesis|data). What makes it effective is that it can harness the capabilities of the planning engine – a core and heavily studied topic in AI– by treating it as a black box that provides p(action|utility) across a variety of scenarios. Suppose one applies this model to the chasing with spatial gap study (Csibra et al., 2003), it depends on the planning engine to know that the optimal action for the big chaser would be to detour, and for a small chaser would be to directly pass through. Therefore, according to this model, humans' ability to flexibly attribute mental states based on observed actions in various scenarios stems from our capacity to plan efficient actions within those contexts.

The rationality principle has deep roots in many fields. It is based on utilitarian philosophy (Bentham, 1789) and serves as the default assumption of human nature in economics (Fishburn, 1970; Von Neumann & Morgenstern, 1944). Overall, it offers a commonsense framework for understanding animate agents. Here, we want to point out an implication of this theory on perception that might be counterintuitive and has been somewhat overlooked. The rich social information extracted from the impoverished motion of simple geometric shapes—as demonstrated by the Heider & Simmel displays—may suggest that perceived animacy is akin to a Rorschach test, where people see patterns from randomness. This impression is consistent with studies of anthropomorphism showing that people tend to see human characteristics where there are none (S. E. Guthrie, 1995; S. Guthrie & Porubanova, 2020), such as seeing faces in clouds

(e.g., pareidolia phenomena). The argument is that because detecting an animate agent is so important, our minds become overly sensitive in identifying them—false alarms are acceptable, but missing an agent (such as a tiger) could mean the difference between life and death. However, this perspective overlooks the fact that the movements in Heider & Simmel displays are not random but meticulously crafted by the authors, reflecting extensive common-sense knowledge about how agents act rationally in an environment. As we mentioned before, merely showing self-propelled objects moving randomly does not result in a compelling perceived animacy—people do not over-attribute agency when viewing haphazard motions. In fact, recreating a set of Heider & Simmel-like displays can require great effort and is considered a significant contribution to the field (e.g., Maslan et al., 2015). Therefore, we argue that perceived animacy is not a Rorschach-like phenomenon precisely because of the rationality principle. This principle imposes a stringent criterion on what is interpreted as an animate agent: the agent must take the most rational action within its environment. Violating this principle means that an object will not be perceived as a goal-directed agent, even if it can effectively pursue a goal but does so in a less rational way.

Studies on infants and adults have shown that human visual perception is highly adapted to detecting efficient goal-directed motions. preferentially attend to chasing motion over control motion, with this preference driven by accelerations and attraction (i.e., movement in a 'heat-seeking' manner), rather than by high turning rates. Additionally, the effect sizes of cues to chasing add linearly or in a marginally diminishing fashion, indicating that infants' attention to chasing is guided by its individual features rather than their overall configuration (Frankenhuis et al., 2013). This dissociation between objective efficiency and subjective goal-directedness is

confirmed by studies on the psychophysics of chasing (Gao et al., 2009). In this study, the perception of chasing was manipulated by adjusting chasing subtlety, a variable that controls the randomness injected into a wolf's heat-seeking pursuit of a sheep. The subtlety ranges from 0° (perfect heat-seeking without noise) to 180° (completely random motion). The results showed that perceived chasing is surprisingly stringent: only chasing with small subtleties (0°–30°) was readily detected by participants. When the subtlety value reached 60° and above, the perception of chasing dropped dramatically, approaching the chance level. These results imply that wolves with moderate chasing subtleties could effectively stalk the sheep without being noticed by subtly deviating from the most efficient chasing path—essentially "hacking" the rationality principle. This is, in fact, a strategy one might employ when trying to stalk others. This finding was confirmed by a "Don't-Get-Caught" task, in which participants controlled the sheep themselves to avoid a hidden wolf among identically looking distractors. The results showed that wolves with moderate chasing subtleties were the most dangerous—they could effectively approach participants while avoiding detection. A similar stalking effect was observed when chasing was disrupted not by spatial noise but by temporally mixing chasing with periodic random motion, suggesting that this phenomenon is not limited to a specific type of noise but is a general effect due to deviation from the most effective chasing strategy as assumed by the rationality principle (Gao & Scholl, 2011). These findings highlight that while an agent may be objectively effective in pursuing a goal, it must act in the most rational way to be perceived as goal-directed by observers.

We have seen that the rationality principle, formulated as an inverse planning model, is a framework supported by both developmental and perceptual studies. Could the perception of the

Heider & Simmel displays be fully explained by employing increasingly powerful black-box planning engines within a Bayesian inference framework? We believe this is unlikely due to the major limitation of the planning engine, namely that it can only work with a predefined utility function. The standard model of AI assumes that defining an agent's utility function is straightforward, and the challenging part is finding the optimal policy to maximize this utility—a task that can be delegated to a black-box planning engine. However, we argue that this view is too optimistic, especially when applied to cognitive science.

Uncovering the intricate structure of the utility functions humans use to explain observed actions, rather than focusing solely on the rational planning process, could be a significant challenge that deserves research attention. Here, we briefly review two lines of work aimed at discovering the utility functions humans adopt when interpreting others' actions.

1) Positive reward vs. negative cost.

   The naive utility calculus proposes that when infants observe actions, they intuitively interpret them based on two types of utilities: positive rewards, representing what the agent aims to achieve, and negative costs, representing what the agent seeks to avoid (Jara-Ettinger et al., 2016). The interplay of these utilities, each varying in magnitude, creates a spectrum of scenarios for social inference. For instance, one study demonstrates how 10-month-old infants use the action costs an agent is willing to incur for reaching an object to infer the value the agent places on that object (Liu et al., 2017). The physical

obstacles the agent needed to overcome were manipulated, including high barriers, wide gaps, or steep slopes. When an agent surmounted significant challenges, infants inferred that the object held high value for the agent. Conversely, if the agent only tackled minor obstacles like shorter barriers, narrower gaps, or gentler slopes and gave up when faced with greater difficulty, infants inferred the object as having lower value to the agent. Similarly, the naive utility calculus extends to how two-year-olds assess agents based on their ability and willingness to help, with the cost of helping manipulated through the agent's competence (Jara-Ettinger et al., 2015). Competent puppets, able to help at a low cost, were judged less nice when they refused to help, while less competent puppets, who incurred higher costs to help, were seen as nicer. This shows that children factor both the cost of helping and willingness when making social evaluations.

2) Goal object as a prioritized type of utility.

In a landmark study, Garcia & Koelling (1966) demonstrated that rats are biologically predisposed to associate sickness with food taste rather than with audiovisual stimuli like lights or sounds. This suggests that not all types of stimuli are equally associable with certain outcomes. Similarly, when humans attribute mental states to observed actions, they do not treat all possible utility functions equally. There is converging evidence showing that reaching an object as a goal holds a prioritized status in human perception, demonstrated in three distinctive ways.

First, Object vs. Location: When 6-month-old infants observe a person reaching for an object in a location, they interpret the goal of the reach as obtaining the object, not

merely moving to a specific location. If the object changes location, infants expect the person to reach for the same object now appearing in a new location rather than reaching for a new object in the old location (Woodward, 1998). This cannot be explained solely by applying the rationality principle, as both the object and the location could equally serve as the utility driving the action.

Second, Goal vs. Source: In linguist theories of verb syntax, the goal or source of an object's change of location is typically encoded by an oblique object, introduced by a preposition. For example, in "She walked to the park," the oblique object "to the park" encodes the goal, while in "He took the book from the shelf," the oblique object "from the shelf" encodes the source. In reality, an agent's motion can simultaneously involve moving toward a goal and away from a source—for example, a person leaving their home (source) to go to the park (goal). In such motions, both reaching the goal and leaving the source could equally be the utility driving the action. Yet, both adults and children are more sensitive to detecting changes in the goal than changes in the source in movie clips (Lakusta & Landau, 2012).

Third, Object vs. Part: While the rationality principle does not impose constraints on what can serve as a goal, perception does. Research has shown that perception assumes the goal an agent is reaching toward must be an individual object (van Buren et al., 2017). When the individuality of the goal is disrupted—for instance, by connecting it to another moving object with a line—the perception of chasing drops dramatically. This demonstrates that humans readily perceive one agent chasing another object but have difficulty perceiving an agent chasing an endpoint of a spatially extended object.

These findings suggest that humans prioritize certain types of utility functions—specifically, those involving discrete objects as goals.

We have reviewed evidence showing that uncovering the intricate structure of human utility functions is a complex task that demands perspectives beyond the rationality principle, which only applies once the utility function is clearly defined. Now, we shift our focus to the planning engine itself. While using AI planning engines—typically modeled as Markov decision processes and solved by reinforcement learning—has yielded productive outcomes (Ho et al., 2021; Jiang et al., 2021; Kleiman-Weiner et al., 2016; Shu et al., 2021) no existing off-the-shelf system can simulate plans with the level of complexity seen in the Heider & Simmel displays. These displays involve long sequences of actions best captured through the composition of verbs (e.g., "hiding," "chasing," "deceiving," "hesitating"), reflecting intricate human intentions. Current planning engines lack representations as rich as verb semantics to compactly describe such human plans. Building a planning engine capable of achieving the complex plans seen in the Heider & Simmel displays may first require understanding how intentional actions are represented in the mind—that is, uncovering how action verb semantics are structured in human cognition. Mainstream AI planning approaches—such as reinforcement learning (inspired partly by animal learning through trial and error) and planning domain definition languages (PDDL, McDermott, 2000)—do not correspond well with the formulations of verb semantics developed in cognitive science. These approaches lack the richness and nuance required to model the sophisticated intentional actions we readily perceive. Therefore, instead of relying on increasingly powerful black-box planning engines, we may need to reconsider our approach. In

the next section, we will explore insights from linguistics to learn what verb semantics can reveal about human intentional actions.

2.5 Ladder 4: Intentional Agent

We started this chapter by clarifying that our focus is on a theory of animate agency as conceptualized by the human mind, not on the scientific theory of agency itself. This distinction is evident when discussing rapid motion and self-propelled motion, which concentrate on the motion itself rather than interpreting it as a response to observations controlled by an agent's mental states—the emphasis of scientific theories of agency. However, as we climb the ladder and reach the utility maximizing agent, we see the boundary between these two perspectives begin to blur. In this context, the model of agency adopted by human perception closely matches the models of agency advocated by decision theory and AI.

The question is whether the utility maximization model is sufficiently rich to capture the complexity of agency in the human mind. Some in the AI community believe it is, as suggested by the title of a recent review article, "Reward is Enough" (Silver et al., 2021). However, from a cognitive science perspective, we argue that reward or expected utility alone is not sufficient. In this section, we will delve into models of agency primarily based on analyses of the semantics of human language. It's important to note that the theories we review are typically considered scientific models of agency, not theories of how animacy is perceived. However, we believe these semantic analyses are highly relevant to perceived animacy, based on two assumptions: (1)

linguistic semantics can be broken down into non-linguistic concepts that are outputs of perception, and (2) language is often learned and used to describe what we observe—just as when interpreting Heider & Simmel displays. Therefore, insights from these semantic theories should apply equally well to our understanding of perceived animacy.

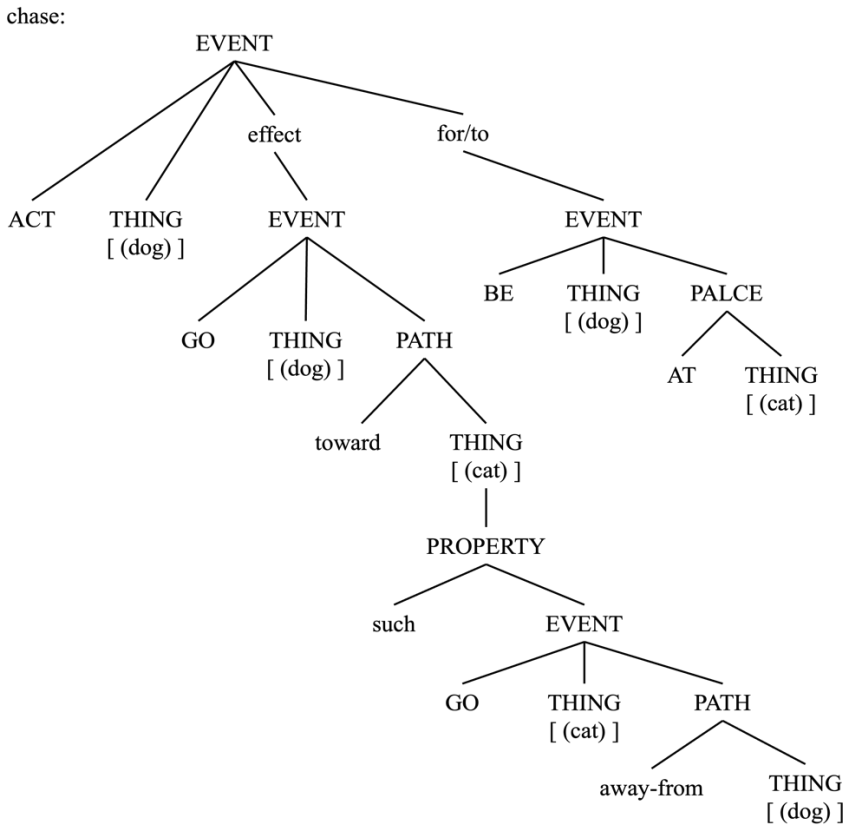
We would like to begin discussing the limitations of utility maximization by referencing plots from Game of Thrones. In a telling conversation, Littlefinger advises Sansa Stark: “Every possible series of events is happening all at once. Live that way and nothing will surprise you.” This is actually a very good description of what a purely utility maximization agent would do: consider all possible futures, take actions that maximize the averaged utility of these futures, but refrain from committing to any one of them. The reason this 'wisdom' needs to be deliberately instructed to the naïve Sansa is exactly because it goes against how humans intuitively process social interactions in daily life—we don’t sum up all the possibilities of future events, indicating that being an agent is more than maximizing expected utility.

One such mental representation outside utility calculus is intention, which is a deliberate mental state that regulates conflicting desires, and commits the agent to a course of action to bring out a fixed future. The distinction between desire satisfaction as utility calculus and achieving an intention is well illustrated by an example: Bill intends to murder his uncle, but while driving and preoccupied with planning the murder, he accidentally runs over a pedestrian who happens to be his uncle (Searle, 1983). Although Bill’s intention indirectly led to his uncle’s death and the outcome aligns with his desire, it is not true that he fulfilled his intention to kill his uncle, as

fulfilling an intention requires a committed sequence of deliberate actions that directly bring about the desired outcome.

This aspect of agency—particularly the role of intention—remains one of the least explored in studies using Heider & Simmel displays. This may help explain why concepts like 'hiding,' 'fleeing,' 'fighting,' and 'breaking' have not been fully grounded in a visual display, as capturing the agent's intentions is essential to understanding the semantic depth of these actions. To illustrate the intricate mental representations of verbs beyond utility maximization, we refer to the semantic analysis of the verb "chase" in the sentence "The cat is chasing the mouse" (Pinker, 1989). When we started working on the psychophysics of chasing, little did we know that the semantics of chasing was notoriously difficult to analyze in linguistics! The semantic breakdown of "chase" here illustrates how *Intention* functions as a distinct concept alongside others like *Event*, *Path*, and *Place* in the conceptual space to represent verb meaning (Fig. 1). "Chase" as an event is composed of two sub-events: one where the cat's goal is to be at the location of the mouse (*Place*), and the other where the cat seeks to achieve this goal by moving along a *Path* toward the mouse, who is simultaneously moving along a *Path* away from the cat. This example highlights how intention interplays with other spatial and event concepts to convey the full meaning of a verb.

Figure 1. The parse tree of the semantics of "chasing"



Note. Illustration of the decomposition of a verb into its core components: intention, event, path, and place. Inspired by Learnability and Cognition: The acquisition of argument structure, by Pinker, 1989, MIT Press.

Linguistic studies have demonstrated that the unique properties of intention are deeply embedded in verb semantics and can influence syntax in various ways. Consider the causative construction, which involves a syntactic structure where an agent (the doer of the action) uses a verb to bring about a change in a patient (the receiver of the action). Crucially, the causative construction is highly sensitive to whether the change in the patient's state is the intended outcome of the action

or merely a means to achieve that outcome (Pinker, 2007). For instance, the verb "to butter" means "to cause butter to be on" something. When a chef spreads butter on a slice of bread, we say, "The chef buttered the bread." However, if the chef puts butter on a knife as a step toward buttering the bread, we wouldn't say, "The chef buttered the knife," because in this case, the knife is merely a means to achieve the intention, not the intended endpoint. More strikingly, causative constructions are not only limited to situations where the outcome is intended but also require that the outcome be a direct result of the agent's intentional action without any intermediate step (Wolff, 2003). For example, participants say a woman "dimmed the lights" when she slid a dimmer switch—not when she turned on her toaster (thereby causing the lights to be dimmed); a man "waved the flag" when he shook a flagpole—not when he raised the flag on a windy day. Note that the distinction between direct and indirect causation is precisely why Bill did not murder his uncle in the car accident, even when the accident was caused by his preoccupation with planning the murder. In other words, such a subtle distinction is not only raised from philosophical reflection; it is so deeply rooted in human intuition that it is directly encoded in the syntax of language.

The limitations of causative verbs on direct causation reveal insights into how intentions are conceptualized beyond mere utility calculus. For instance, while you can 'ring a bell,' you cannot 'cry a baby' or 'laugh a friend' because actions like crying or laughing are seen as under voluntary control, even if external events, like a joke, prompt the reaction. This linguistic nuance reflects the philosophical distinction between actions caused by external forces and those arising from internal intent (Levin, 1985; Pinker, 2007). Searle (2001) notes that, under utility maximization, an agent's actions are driven directly by desires, leaving no gap for free choice—like a cocaine

addict frantically searching for a fix, compelled by a powerful craving rather than true choice. While this behavior aligns with utility maximization, it lacks freedom and rationality. Searle argues that what distinguishes human rationality is precisely the psychological gap between desires and actions, allowing for intentional deliberation and the freedom to choose among competing desires. Causative verbs and their limitations reflect this assumption, where true agency involves the freedom to navigate conflicting impulses and decide one's path.

The intentional nature of agency beyond utility maximization has recently been explored in studies on both children and adults (Cheng et al., 2023; Chu & Schulz, 2022). In one study inspired by the thought experiment Buridan's Ass, where a donkey struggles to choose between two equal options (e.g., water and food at an equal distance), adult participants navigated between two equally desirable destinations. They showed strong commitment to their initial choice, even when external forces pushed the participant-controlled off its course, making an alternative goal more optimal. This suggests that once a decision is made, humans stick to it despite changing circumstances. Interestingly, this commitment forms the basis for intention-based social coordination. Even when explicitly told that the task was individual, participants coordinated their actions by signaling their intentions and interpreting those of others. The inverse planning model revealed that participants delayed revealing their intended destination compared to when they played alone. The participant who revealed their intention later overwhelmingly chose a different destination to avoid the one earlier indicated by the other player, suggesting a phenomenon of intention-based ownership—where clearly signaling one's intent to reach a goal leads others to perceive it as claimed, prompting them to pursue alternative options. This study highlights that humans rely on the demonstration and perception of intention to navigate social

interactions, even without explicit coordination. While the Heider & Simmel displays offer rich social information, they provide only a passive experience. In real life, humans use social perception not just to understand others but to engage and interact with them. This study marks a step toward understanding how social perception operates in active, real-world contexts.

3 Open questions

In this chapter, we are guided with the following assumptions from cognitive semantics: a) agency is a basic type of concept that enters conceptual space, along with other concepts like object, place, and event; b) it is a concept that emerges early in young infants as a type of core knowledge, and continues to support and constrain adults' perception; c) it serves as the building block of linguistic semantics, used as a basic ingredient in defining the semantics of many verbs.. Under this framework, we review empirical evidence from the three fields (perception, development, and linguistics), arranging them into a ladder of agency according to the complexity of mental representations attributed to the observed motion, ranging from rapid motion to intentional commitment. The goal is not only to find parallels between perceptual, developmental, and linguistic studies, as shown in recent research on concepts such as object (Scholl & Pylyshyn, 1999; Soja et al., 1991; VanMarle & Wynn, 2011), event boundary (Baldwin et al., 2001; Papafragou & Ji, 2023; Strickland et al., 2015), and symmetry (Hafri et al., 2023), but also to recognize the gaps that can hopefully motivate future research.

Here are a few open questions we noticed when reviewing studies from the three fields.

1) There is an uneven distribution of evidence across different levels of animacy. One notable pattern is that evidence from different fields is unevenly distributed across the agency ladder. At the most basic level—distinguishing animals from plants—we observe converging evidence from all three fields. This convergence underscores a shared understanding across disciplines regarding the basic definition of animate agent: the ability to act rapidly to change the environment. However, as we move up to the intermediate levels involving self-propelled motion and utility-maximizing actions, the evidence becomes less balanced. Here, we find strong support from perceptual and developmental studies, but linguistic evidence is notably lacking. To the best of our knowledge, there are no syntactic structures in language that explicitly mark the distinctions between self-propelled and inert motion or between utility-maximizing and non-utility-maximizing actions. A significant shift occurs when we ascend to the highest levels of this hierarchy. Distinctions such as intention vs. desire, means vs. ends, direct vs. indirect causality, and forced vs. voluntary choice come into focus. At this level, linguistic evidence richly supports these distinctions, while evidence from perceptual and developmental studies is only beginning to surface. One possible explanation for this uneven distribution of evidence is that self-propelled, and utility-maximizing motions are part of the mechanisms within the perceptual module that generate the concept of agency as an output in the conceptual space. These internal perceptual processes are not part of the conceptual space and therefore cannot be directly reflected in linguistic semantics—much like the mechanisms by which we perceive 3D depth, which cannot be consciously experienced or directly encoded in language. In contrast, intention—with all the characteristics that distinguish it from desires—is clearly represented in the output of

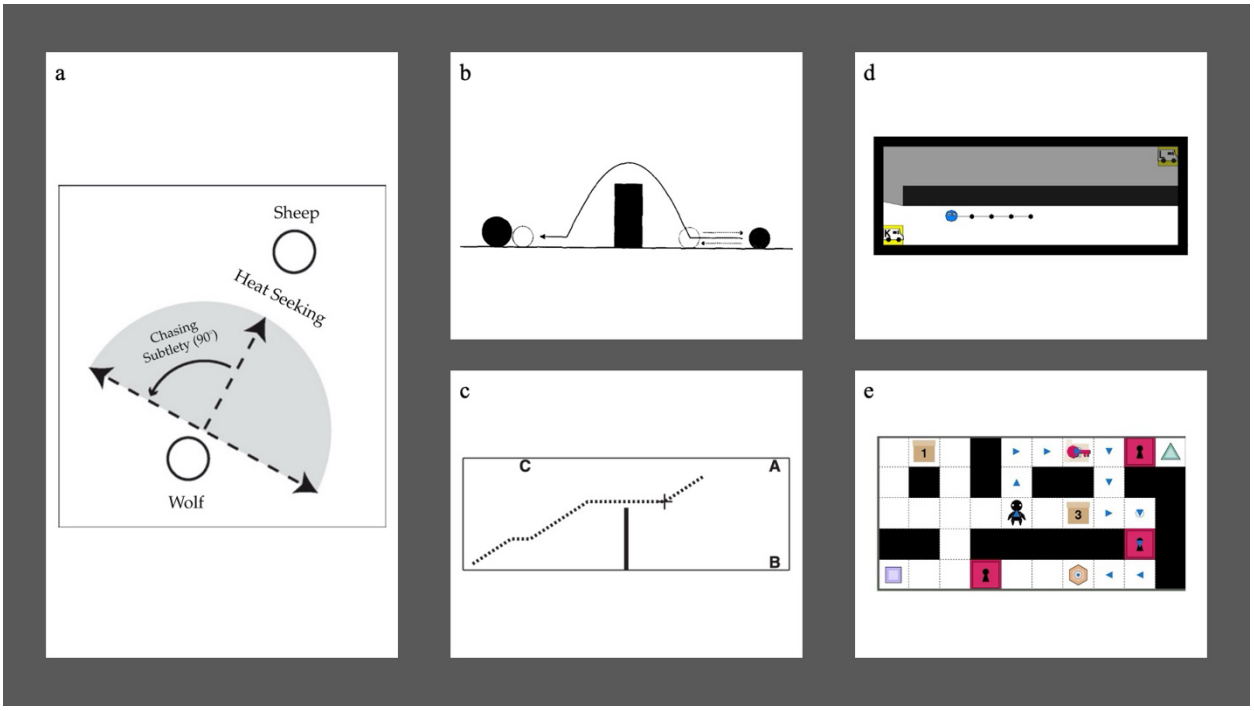
perception, allowing its properties to be revealed through linguistics. What remains to be understood is the perceptual mechanism that goes beyond self-propelled motion and utility calculation to support the complex qualities of intention that semantics requires.

2) While we highlight the parallels between perception, infant core knowledge, and cognitive semantics, there are certainly tensions among them. Starting with perception and core knowledge, theorists have emphasized that core knowledge goes beyond mere perception and constitutes a form of core cognition (Carey, 2009). For a detailed discussion on this topic from a perception perspective, see Scholl (2024). Instead of getting into the broad debate of perception vs. cognition, here, we want to point out one research question raised in reviewing studies of the Heider & Simmel displays. The Bayesian inverse planning model has been applied to explain a broad range of tasks (Fig. 2), including modeling chasing as heat-seeking (Gao et al., 2019), inferring utility under the rationality principle (Gergely et al., 1995; Liu et al., 2017), identifying goal pursuit through obstacle avoidance (Baker et al., 2009), inferring a hierarchy of desires by jointly analyzing desires and beliefs (Baker et al., 2017), and attributing an agent's ultimate goal through instrumental reasoning about hidden objects as sub-goals (Ying et al., 2024). With such a broad range of applications, it is reasonable to assume that the boundary between perception and cognition lies somewhere along this spectrum. Future research should clarify the limits of the rationality principle in perception: Is perception restricted to certain fixed cues that may be hard-coded in the visual system, or does it have the flexibility to assess the rationality of actions by dynamically applying utility calculus to

environmental constraints? Or, alternatively, does this more adaptive analysis rely on a cognitive inference process?

Figure 2

Examples of Tasks Applying the Bayesian inverse planning Model



Note. a) Chasing identification, where individuals detect the chasing behavior and identify the chaser and the one being chased based on motion that deviates from the heat-seeking manner to varying degrees (with possible deviations ranging from 0 to 180 degrees, 90 degrees deviation shown in the figure); Reprinted with permission from *The cognitive architecture of perceived animacy: Intention, attention, and memory* by Gao et al., 2019, *Cognitive Science,* Wiley. b) Utility inference under the rationality principle, where humans assume the agent will take the shortest path constrained by the barrier; Reprinted with permission from *Taking the intentional*

stance at 12 months of age by Gergely et al., 1995, *Cognition,* Elsevier. c) Goal inference task, where participants infer the goal (A, B, or C in the figure) based on the observed actions through inverse planning processes; Reprinted with permission from *Action understanding as inverse planning* by Baker et al., 2009, *Cognition,* Elsevier. d) Joint desire and belief inference, where individuals infer both desires (the preference for the food truck—whether it is K in the left coner, L behind the wall, or M, which hasn't appeared) and beliefs (whether the food truck exists behind the black wall, and if so, which one) of the agents based on their actions. Used with permission from *Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution* by Baker et al., 2011*, Proceedings of the Annual Meeting of the Cognitive Science Society, licensed under CC BY 4.0*. e) Coherent sets of goals, beliefs, and plans joint inference, where participants jointly infer goals (wanting the hexagon), beliefs (believing that the key for the lock on the path to the hexagon is in box 2), and plans (first going to box 2 to get the key, then unlocking the lock, and finally reaching the hexagon, while planning the shortest path in sequence) using a Bayesian Theory-of-Mind (BToM) framework. Used with permission from *Grounding Language about Belief in a Bayesian Theory-of-Mind* by Ying et al., 2024, *ArXiv Preprint, licensed under CC BY 4.0*.

3) The tension between perception and semantics also becomes apparent when examining how causal interactions are modeled. Semantic analysis reveals that intentional actions often hinge on the causal relationship between an agent and a patient. Additionally, our discussion of self-propelled motion shows that a model of agency relies on an underlying physical model that simulates movement based on self-exerted force. However, a fundamental difference exists in how causality and physical interactions are

conceptualized in perceptual versus linguistic studies. Recent perceptual studies suggest that interactions between rigid objects can be approximated by a Newtonian physics engine (Battaglia et al., 2013; Ullman et al., 2017). In contrast, the force-dynamic model based on linguistic analysis of causal verbs (Talmy, 1988) departs from Newtonian principles in several ways. First, they introduce two distinct roles: the agonist, an entity with an intrinsic tendency, and the antagonist, which either supports or opposes this tendency—roles that have no counterpart in Newtonian physics. Second, the intrinsic tendency of the agonist resembles the Aristotelian dynamics , which predated Newtonian physics in the history of science. In Aristotelian dynamics, motion is driven by an object's intrinsic nature and external forces, with objects striving to return to their natural state, a concept that resonates with Talmy's idea of an intrinsic tendency that shapes the dynamics of agents. Third, the interaction between agonist and antagonist is unidirectional, with the antagonist acting upon the agonist without a reciprocal counterforce, contradicting Newton's principle of equal and opposite forces. Which type of causal-physical model is used in perceived animacy? In studies on perceived chasing, the wolf (agonist) was restrained by a master (antagonist) using a leash, causing the wolf's movement to significantly deviate from a direct, heat-seeking path (Tang et al., 2021, 2024). Results showed that participants readily detected chasing despite the large spatial deviations, provided that the leash was modeled as a one-way impact of the master on the wolf, as in Talmy's framework (1988). When the interaction was instead modeled symmetrically in a realistic Newtonian physics engine, perceived chasing dropped significantly. One possible explanation is that in human perception, approximate Newtonian physics applies mainly to inert objects, while animate, self-propelled motion

requires a different model—better captured by the semantics of causal verbs. This hypothesis needs further testing across a broader range of social and physical interactions.

4) Finally, we acknowledge that our definition of animate agency remains incomplete, even at the highest level of our agency ladder. One conspicuously missing component is emotion. A significant part of the experience of watching the Heider & Simmel displays involves interpreting emotional states—the small triangle appears frightened, and the large triangle seems furious when it loses its target. While emotions are well-studied in psychology, they are rarely incorporated into models of agency. From a utility-maximization perspective, it's unclear why a rational agent would need emotions—after all, we wouldn't want autonomous cars equipped with emotions that might lead to road rage. Yet emotions undeniably drive human actions, sometimes enriching or even overriding strict rationality in our decisions. Future research needs to explore how emotion might be integrated into agency frameworks in a way that complements rational models of agency.

4 Conclusion

This chapter has explored the cognitive architecture and mechanisms underlying perceived animacy, organizing empirical findings from perceptual, developmental, and linguistic studies into a "ladder of animacy." Overall, the findings support that animate agency is a foundational

concept within the conceptual space—perceived spontaneously, developed early in infancy, and central to verb semantics. However, at different levels, distinct emphases emerge: perceptual and developmental evidence primarily highlights the self-propulsion and utility-maximizing nature of animate motions, while linguistic semantics centers on the intentional aspects of actions, such as commitment, direct agent-patient interactions, and volitional freedom. These differences suggest that future research should explore how these intentional properties of agency are represented within perceptual processing, providing a foundation upon which verb semantics can be built.

Acknowledgments

This work was supported by ONR Vision Language Integration grant (PO# 951147:1) to TG.

Reference

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bentham, J. (1789). From an introduction to the principles of morals and legislation. In *Literature and Philosophy in Nineteenth Century British Culture*. Clarendon Press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Cheng, S., Zhao, M., Tang, N., Zhao, Y., Zhou, J., Shen, M., & Gao, T. (2023). Intention beyond desire: Spontaneous intentional commitment regulates conflicting desires. *Cognition*, *238*, 105513.

Chu, J., & Schulz, L. (2022). " Because I want to": Valuing goals for their own sake. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 1263–1269).

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Dennett, D. C. (1996). *Kinds of minds: Towards an understanding of consciousness*. Phoenix.

Di Giorgio, E., Lunghi, M., Simion, F., & Vallortigara, G. (2017). Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental Science*, *20*(4), e12394.

Di Giorgio, E., Lunghi, M., Vallortigara, G., & Simion, F. (2021). Newborns' sensitivity to speed changes as a building block for animacy perception. *Scientific Reports*, *11*(1), 542.

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253–268.

Fishburn, P. C. (1970). *Utility theory for decision making*. John Wiley and Sons.

Frankenhuis, W. E., House, B., Barrett, H. C., & Johnson, S. P. (2013). Infants' perception of chasing. *Cognition*, *126*(2), 224–233.

Gao, T., Baker, C. L., Tang, N., Xu, H., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive Science*, *43*(8), e12775.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing : A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154–179. https://doi.org/10.1016/j.cogpsych.2009.03.001

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*, 123–124.

Gelman, R., Durgin, F. H., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*. Clarendon Press.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*(4), 480–487.

Guthrie, S. E. (1995). *Faces in the clouds: A new theory of religion*. Oxford University Press.

Guthrie, S. E., & Porubanova, M. (2020). Faces in clouds and voices in wind: Anthropomorphism in religion and human cognition. *Routledge Handbook of Evolutionary Approaches to Religion. Routledge*.

Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2023). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology: General*, *152*(2), 509.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*(11), 2246.

Jackendoff, R. S. (1985). *Semantics and cognition* (Vol. 8). MIT Press.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640.

Jiang, K., Stacy, S., Wei, C., Chan, A., Rossano, F., Zhu, Y., & Gao, T. (2021). Individual vs. joint perception: a pragmatic model of pointing as communicative smithian helping. *ArXiv Preprint ArXiv:2106.02003*.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1679–1684). Cognitive Science Society.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517–544.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Levin, B. (1985). Introduction. In B. Levin (Ed.), *Lexical Semantics in Review (Lexicon Project Working Papers # 1)*. MIT Center for Cognitive Science.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science*, *16*(8), 601–608.
Luo, Y., Kaufman, L., & Baillargeon, R. (2009). Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology*, *58*(4), 441–486.
Malik, M., & Isik, L. (2022). *Relational visual information explains human social inference: a graph neural network model for social interaction recognition*.
Maslan, N., Roemmele, M., & Gordon, A. S. (2015). One hundred challenge problems for logical formalizations of commonsense psychology. *2015 AAAI Spring Symposium Series*.
McDermott, D. M. (2000). The 1998 AI planning systems competition. *AI Magazine*, *21*(2), 35.
Michotte, A. E. (1963). *The perception of causality (TR Miles, Trans.)* (T. Miles, Trans.; Vol. 2). Methuen & Co.
Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.
Opfer, J. E., & Gelman, S. A. (2011). Development of the animate-inanimate distinction. *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, *2*, 213–238.
Papafragou, A., & Ji, Y. (2023). Events and objects are similar cognitive entities. *Cognitive Psychology*, *143*, 101573.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. The MIT Press.
Pinker, S. (1997). How the mind works. *WW Norton&Company*.
Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. Penguin Books.
Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, *1*(2), 56–61.
Poulin-Dubois, D., Lepage, A., & Ferland, D. (1996). Infants' concept of animacy. *Cognitive Development*, *11*(1), 19–36.
Pratt, J., Radulescu, P. V, Guo, R. M., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science*, *21*(11), 1724–1730.
Russell, S. (2019). *Human compatible: AI and the problem of control*. Viking Press.
Russell, S., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.
Scholl, B. J. (2024). Perceptual (roots of) core knowledge. *Behavioral and Brain Sciences*, *47*, e140.
Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*(2), 259–290.
Searle, J. R. (1983). *Intentionality: an essay in the philosophy of mind*. Cambridge University Press.
Searle, J. R. (2001). *Rationality in action*. MIT press.
Shu, T., Peng, Y., Zhu, S.-C., & Lu, H. (2021). A unified psychological space for human perception of physical and social events. *Cognitive Psychology*, *128*, 101398.
Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.
Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, *38*(2), 179–211.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.
Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.
Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, *112*(19), 5968–5973.
Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.
Tang, N., Gong, S., Liao, Z., Xu, H., Zhou, J., Shen, M., & Gao, T. (2021). Jointly Perceiving Physics and Mind: Motion, Force and Intention. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 735–741). Cognitive Science Society.
Tang, N., Xu, E., Gong, S., Zhou, J., Shen, M., & Gao, T. (2024). Perceiving animacy through schematic intuitive physics: Shared conceptual structure of animacy between vision and language. *Journal of Vision*, *24*(10), 803.
Todorov, E., Erez, T., & Tassa, Y. (2012). Mujoco: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033.
Träuble, B., Pauen, S., & Poulin-Dubois, D. (2014). Speed and direction changes induce the perception of animacy in 7-month-old infants. *Frontiers in Psychology*, *5*, 1141.
Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943–951.
Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Psychophysics*, *68*, 1047–1058.
Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.
van Buren, B., Gao, T., & Scholl, B. J. (2017). What are the underlying units of perceived animacy? Chasing detection is intrinsically object-based. *Psychonomic Bulletin & Review*, *24*, 1604–1610.
VanMarle, K., & Wynn, K. (2011). Tracking and quantifying objects and non-cohesive substances. *Developmental Science*, *14*(3), 502–515.
Von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. In *(No Title)*. Princeton University Press.
Westfall, M. (2023). Perceiving agency. *Mind & Language*, *38*(3), 847–865.
Wick, F. A., Alaoui Soce, A., Garg, S., Grace, R. C., & Wolfe, J. M. (2019). Perception in dynamic scenes: What is your Heider capacity? *Journal of Experimental Psychology: General*, *148*(2), 252.
Wikipedia contributors. (n.d.). *Animacy*. Wikipedia. Retrieved April 9, 2025, from https://en.wikipedia.org/wiki/Animacy
Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.
Yamamoto, M. (2006). *Agency and impersonality: Their linguistic and cultural manifestations*. John Benjamins Publishing Company.
Ying, L., Zhi-Xuan, T., Wong, L., Mansinghka, V., & Tenenbaum, J. (2024). Grounding Language about Belief in a Bayesian Theory-of-Mind. *ArXiv Preprint ArXiv:2402.10416*.