Beyond the Courtroom:
Agency and the Perception of Free will

Edouard Machery[1*], Markus Kneer[2], Pascale Willemsen[2], Albert Newen[3]

1) Department of History and Philosophy of Science, University of Pittsburgh, United States

2) Institute of Philosophy, University of Zurich

3) Institute for Philosophy II, Ruhr-University Bochum

* Corresponding Author

Abstract

In this paper, we call for a new approach to the psychology of free will attribution. While past research in experimental philosophy and psychology has mostly been focused on reasoning-based judgment ("the courtroom approach"), we argue that like agency and mindedness, free will can also be experienced perceptually ("the perceptual approach"). We further propose a new model of free will attribution—the agency model—according to which the experience of free will is elicited by the perceptual cues that prompt the attribution of agency. Finally, developing new stimuli that fit the perceptual approach, we present some preliminary evidence in support of the agency model.

1. Introduction

Philosophers, psychologists, and neuroscientists have proposed two different families of models of how we ascribe mental states, such as beliefs, desires, action plans, or intentions: the Theory Theory (e.g., Gopnik & Wellman, 1992; Nichols & Stich, 2003) and Simulation Theory (e.g., Goldman, 1992, 2006). Theory Theory Models claim that we ascribe mental states on the basis of a folk-psychological theory, while Simulation Theory models hold that mental state attribution results from the ascriber simulating what kind of mental states they would have, were they in the ascribee's situation. Despite their disagreements, both the Theory Theory and Simulation Theory, at least in their original variants (but see Gallese & Goldman, 1998), are committed to the view that mindreading belongs to higher cognition. According to the Theory Theory, mindreading results from some form of theoretical reasoning, and, according to Simulation Theory, it requires understanding how one's situation differs from the assignee's situation.

However, in the last two decades, cognitive-scientific research has revealed that philosophers, psychologists, and neuroscientists had often overlooked at least one important additional access to understanding others, namely a low-level response to a variety of perceptual, particulary visual, cues. This shift in perspective is mirrored in the discovery of *implicit* false belief understanding (Baillargeon et al., 2010; Southgate et al., 2007), which is measured by the observation of gaze direction and duration as well as anticipatory looking. Furthermore, in philosophy, it led to arguments for the direct perception of some mental phenomena (Carruthers 2015; Gallagher, 2008; Newen et al., 2015), and, in psychology, to the investigation of the perception of goals in adults, children, and babies (e.g., Csibra et al., 1999; Reid, 2007). Perceptual access to mental states appears not to require reasoning; rather, perceptual cues lead to the experience and judgment that the perceived object is an agent with some specific goals.

In this chapter, we argue that a similar shift in perspective is necessary for understanding the attribution of free will. Judgments based on visual cues have not only been overlooked for

understanding the attribution of beliefs, desires, intentions, and so on (what we will call "mindedness"), but also for understanding the attribution of free will. Experimental philosophers working on free will seem to assume that determining whether an agent has free will or performed an action freely is a higher cognitive endeavour. For them, it would seem, we *reason* about free will and control. There has been much theoretical and empirical debate, for instance, about whether or not laypeople consider determinism consistent with free will (e.g., Hannikainen et al., 2019; Murray & Nahmias, 2014; Nahmias et al., 2005; Nichols, 2004; Nichols & Knobe, 2007; Sarkissian et al., 2010). Because people are asked whether some unexpected or unusual conditions would defeat their usual expectation that people are in control of their actions, they are prime to engage in some form of conscious or unconscious reasoning, even if their judgments are influenced by emotions, as some have claimed (Nichols & Knobe, 2007; but see Feltz & Cova, 2014). We will call this way of studying the psychology of free will attribution *"the courtroom approach"* because it treats the attribution of free will as similar to what a judge does when she decides whether an agent had the capacity to control her behavior (Hollander-Blumoff, 2012). We believe it is no accident that experimental-philosophical debates have taken place against a background of concerns about responsibility and punishment in the law and morality.

While the attribution of free will can sometimes involve reasoning, real-life judgments about whether an agent acts freely are often elicited more directly by the perception of the assignee's behavior. In everyday life, we do not merely reason to free will, we *see* it. The first contribution of this chapter is thus to invite experimental philosophers and psychologists to go beyond the courtroom approach and pay greater attention to the *perception* of free will and control in our everyday interactions. Thus, we develop a new paradigm to study free will judgments: the *"perceptual approach."*

What makes us see free will? The second contribution of our chapter is to argue that free will is often attributed on the basis of the very cues that lead to the perception of agency.

When someone is perceived as an agent, we also tend to perceive them as acting freely. Thus, on our view, in many everyday interactive situations, the attribution of free will is primarily a perception-based judgment that is anchored in the perception of agency. We call this model *"the agency model of free will attribution."*

Here is how we will proceed. In Section 2, we contrast our approach ("the perceptual approach") with previous studies of free will in psychology and experimental philosophy, and we propose our new model of free will attribution. In Section 3, we describe the new experimental materials that were developed to bring about the perceptual approach. Section 4 reports our results. Section 5 discusses their significance both for the agency model of free will attribution and for the existing debates about free will judgment in psychology and experimental philosophy, and it also highlights the limitations of our work.

2. The Agency Model of Free Will Attribution

In 1944, Heider and Simmel presented subjects with a short, black-and-white, animated video in which three geometrical figures (a circle and two triangles of different sizes) move across a surface. In addition, the video showed a fixed, large, rectangular object that could be opened and closed. Although the three geometrical figures looked nothing like actual agents (people or animals), the vast majority of participants in this famous study readily perceived them as agents and they described the scene in agentic terms, attributing intentions, desires, and beliefs. The only agency cues participants saw were the movements of the geometrical figures, which seemed purposeful, with the figures changing direction, accelerating, and decelerating, often in response to the other figures' movements. Perception of agency on the basis of this kind of cues is robust across individuals, including children, and cultures (e.g., Barrett et al., 2005; Bowler & Thommen, 2000; Gao et al., 2010), while being highly sensitive to small variations in the stimuli (Gao et al., 2010). The perception of agentic behavior appears to depend largely on three features—directionality, discontinuity, and responsiveness (Santos et al., 2008)—which we will

call "agency cues." The object must not follow a physically determined trajectory. Instead of moving in a straight line, the object changes course without any contact with other physical objects. It should display a discontinuous movement pattern. For instance, it can stop and accelerate. Finally, the object should respond and react to other objects in its environment.

We propose that when we see an entity as an agent, we tend to see it as having a host of other properties. Previous research already suggests that once an entity is perceived as an agent, we are disposed not only to attribute to it mental states such as intentions, desires, and beliefs, but also to see it as conscious (Arico et al., 2011). The agency model of free will attribution hypothesizes that perceptual cues of agency also lead to viewing the entity as not only minded and conscious, but also as free and in control of its own behavior. Thus, the cues leading people to view an entity as an agent also lead us to view it as free and in control of its behavior. When we see the geometrical figures in Heidel and Simmel's experiment as agentive, we also tend to view them as being in control of their behavior and thus as being free to move one way or another. We thus predict that if an entity as simple as a geometrical figure moves in a directional, discontinuous, and responsive manner, it will be perceived as free.

Noticeably, the agency model of free will is third-personal. It grounds free will judgment in the third-personal perception of others' agency. In this respect, it differs from approaches that highlight the first-personal experience of one's own actions. Wegner (2002, 2003), for instance, has related the belief in free will to the conscious experience of will as the source of one's intentional actions.

We acknowledge that the agency model of free will is underspecified in important respects. For one, the relation between the perceptions of agency, mindedness, and free will is not specified. Perceptual cues could, for instance, lead us to perceive an object as an agent, and this perception could incline us to view it as minded and free (Figure 1). Alternatively, agency, mindedness, and free will (as well, perhaps, as consciousness) could be part of a single package, which we tend to perceive in response to some low-level perceptual cues (Figure 2).

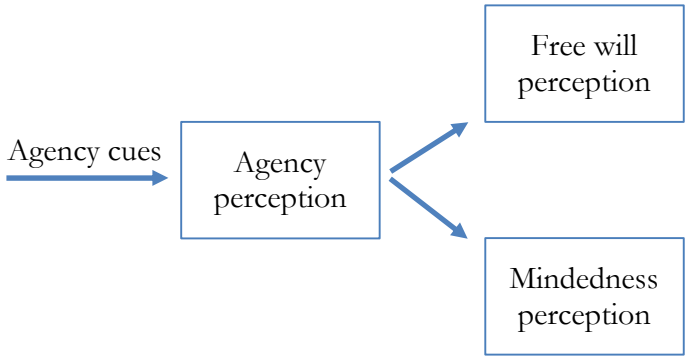

Figure 1: First Version of the Agency Model

Figure 2: Second Version of the Agency Model

In addition, we remain non-committal about the exact nature of the experience of free will (for related debates for the perception of agency, see Scholl & Gao, 2013): Is it genuinely perceptual? If so, in what sense? Is it affected by beliefs, desires, expectations? We believe that we are not yet in a position to address these gaps. Our goal is more modest. We want to loosen the grip of the courtroom approach on the study of free will attribution in experimental philosophy and psychology, developing experimental tools to examine the perception of free will, and we want to show that there is a plausible connection between the perception of agency and of free will.

We would also like to emphasize that the agency model does not deny the relevance of reasoning for the attribution of free will. We do sometimes reason to decide whether someone was in control of her actions. The experimental-philosophy studies treating free will judgments as the output of some form of reasoning might thus be onto something. That said, we aim to demonstrate that there is more to the psychology of free will attribution than reasoning.

3. Designing New Materials for the Perceptual Approach

As noted in the introduction, one of the goals of this chapter is to develop an alternative to the courtroom approach to free will attribution, which acknowledges that many of our judgments about control and free will result from the perceptual experience of free will. Studying the perception of free will requires developing new experimental stimuli. Experimental-philosophy stimuli are typically verbal, describing the behavior of an agent in various situations (e.g., in a deterministic world or in a situation where the agent could not have done otherwise). Shortcomings of the vignettes often used in experimental philosophy studies have already been discussed (Clark et al., 2019; Nadelhoffer et al., 2020; Rose et al., 2017), but our concern here is different: While such stimuli might make sense to study reasoning-based free will judgments, they are inadequate to study the perception of free will since they do not involve any perceptual cues.

To develop new experimental materials, we were inspired by Heider and Simmel's experimental paradigm, which examined the perception of agency and mindedness by presenting participants with simple visual cues embodied by geometrical figures. Similarly, we developed short animated videos displaying the movement of a marble and its interactions with its environment.[1] Agency cues are manipulated across animated videos to determine whether, as predicted by the agency model, participants treat a simple geometrical figure (a marble) as free when it embodies agency cues. In half of the experimental conditions, a simple marble follows a path clearly determined by the layout of the room and its original motion; in the other half, the marble embodies the agency cues. Its movement is clearly not determined by the layout of the room; it is discontinuous; and the marble interacts with its environment (Figures 3 and 4).

[1] All supplementary material, including data files, statistical analyses, and the video stimuli can be found here: https://osf.io/8vdgz/?view_only=98e6b67573b94d3094d1cfc5f2b7ffa3.

We also manipulated the moral valence of the interactions of the marble. Experimental-philosophy studies have shown that moral valence influences judgments of various kinds (e.g., Newman et al., 2014 on true self; Hitchcock & Knobe, 2009; Willemsen & Kirfel, 2019, and Sytsma, 2020 on causation), including attribution of intentionality (e.g., Knobe, 2003), conative states (Tannenbaum et al., 2007), and doxastic states (e.g., Beebe & Buckwalter, 2010; Kneer 2018, 2021; Kneer et al., in press). What's more, using verbal stimuli, psychologists have found that people ascribe more free will for morally bad actions than morally good actions (e.g., Clark et al., 2014; Monroe & Ysidron, 2021). On the other hand, Danks et al. (2014) have provided evidence that at least for causal judgments, the influence of moral valence might be limited to verbal stimuli and be absent when the stimuli are visual (but see Gerstenberg & Icard, 2020). In half of the experimental conditions, the marble enters the scene and allows a character (a mouse) to bring about its goal; in the other half, the marble prevents the mouse from bringing about its goal. The marble fosters or hinders the mouse's goal by following either a purely physical path or an animate movement pattern.

Thus, in a fully factorial 2×2 design (Agency: Agent vs. Object; Valence: Help vs. Hinder), participants were randomly presented with one of four animated videos. In the four videos, a mouse is trying to open a door to enter into the neighboring room, which contains cheese. In the two Help conditions, the marble opens the door; in the two Hinder conditions, it prevents the mouse from opening the door. In the two Object conditions, the marble follows a simple physical trajectory that obeys the mechanical laws of motion. In the two Agent conditions, the marble's movement is self-propelled and discontinuous, and the marble interacts with the mouse.

Given these stimuli, the agency model of free will attribution makes the following predictions:

(i) Participants in the Agent conditions will be more likely to treat the marble as free and in control of its actions than in the Object condition.

(ii) Free will judgments do not depend on participants' judgment that the movement of the marble is determined.

Prediction (ii) follows from the fact that the agency model of free will attribution singles out simple visual cues (viz. the agency cues) and thus does not take determinism to be relevant for the experience of free will. Prediction (ii) stands in contrast with the usual practice in experimental philosophy that connects lay people's understanding of free will with issues related to determinism.

We also predicted that we would observe an effect of the Agency variable on the attribution of mindedness:

(iii) Participants in the Agent conditions will be more likely to assign doxastic states (beliefs and knowledge) and conative states (desires and intentions) than in the Object condition.

The agency model of free will attribution does not make any prediction about the impact of moral valence on free will judgments and existing work is inconclusive about it as well. This part of the study is thus exploratory.

(iv) The existing literature on the impact of moral valence on judgment suggests that participants should be more likely to treat the agent as having doxastic states (Beebe & Buckwalter, 2010), conative states (Knobe, 2003), and possibly free will (Clark et al., 2014) in the Hinder condition than in the Help condition. On the other hand, moral valence may not impact judgment when the stimuli are visual rather than verbal (Danks et al., 2014).

4. Experiment

4.1 Participants

We recruited 813 participants on Prolific. In line with the preregistered criteria, 31 participants who were not native English speakers or failed an attention check were excluded, leaving a sample of 782 participants (female: 64%; age M=39 years, SD=13 years, range: 18–83 years).[2]

4.2 Methods and Materials

The experiment followed a 2 (Agency: Agent vs. Object) × 2 (Valence: Help vs. Hinder) between-subjects fully crossed factorial design, and participants were randomly assigned to one of the four conditions. In all conditions, participants saw a 12-seconds movie. The four movies showed the marble entering one of the rooms, A or B, via the red holes (Figures 3 and 4). Depending on the Agency condition, it followed a path corresponding to the law of physics (blue, dotted path indicating the Object condition) or a path clearly violating the law of physics (red, continuous path indicating the Agent condition).

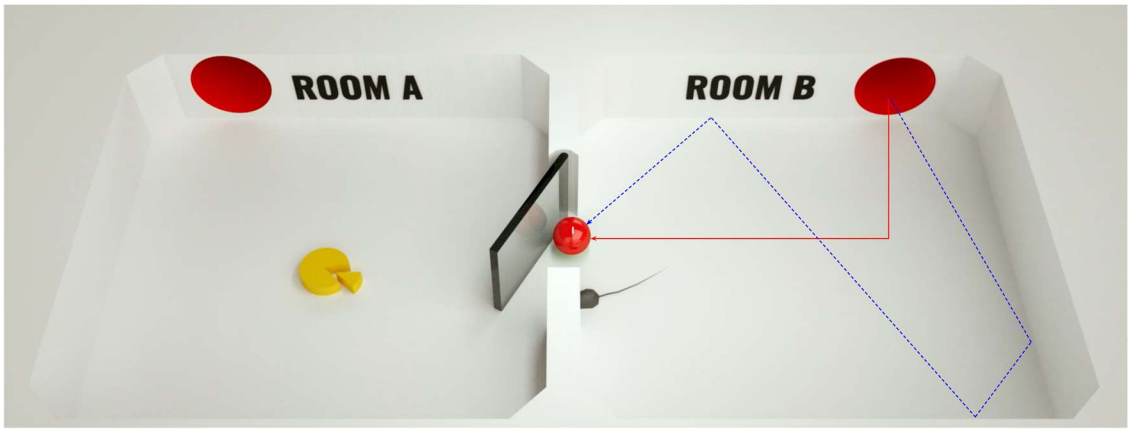

Figure 3: Helping (Red, Continuous Path Indicates Agency and Blue, Dotted Path No Agency).

[2] See aspredicted.org/rf37v.pdf. Because of technical issues we were not able to implement one of the 3 exclusion criteria (duration).

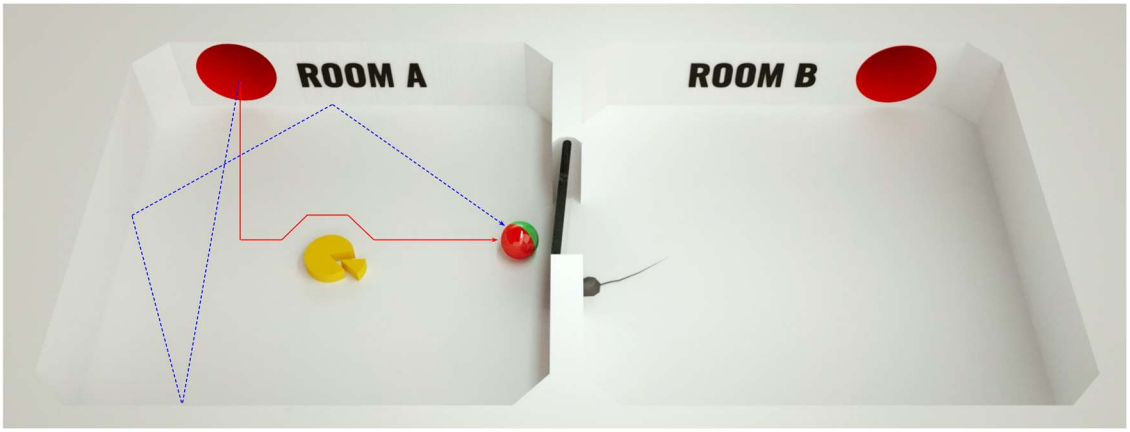

Figure 4: Hindering (Red, Continuous Path Indicates Agency and Blue, Dotted Path No Agency).

Having watched one of the four animated videos, participants read the following ten questions in a list. They then had to watch the video again before answering the same questions in random order. All responses were collected on 7-point Likert scales, anchored at 1=definitely not and 7=definitely yes (unless otherwise indicated).

Free will

Did the marble seem to [open/block] the door of its own free will?

Did the marble seem to have control over its movements?

Doxastic states

Did the marble seem to believe that the mouse wanted to open the door?

Did the marble seem to know that the mouse was interested in the cheese?

Conative states

Did the marble seem to intentionally allow the mouse to enter [prevent the mouse from entering] the other room?

Did the marble seem to want to [open/block] the door?

Moral judgments

Did it seem the case that the marble should be punished or rewarded for [opening/blocking] the door? (anchored at 1= definitely punished and 7= definitely rewarded)

Did the marble seem blameworthy or praiseworthy for [opening/blocking] the door? (anchored at 1= definitely blameworthy and 7= definitely praiseworthy)

Determinism

Did the movement of the marble seem to obey the laws of physical movement?

Did the path of the marble seem to be determined by the way it entered the room and the shape of the room?

4.3 Results

The alpha coefficients for all pairs of dependent variables were larger than .7 (Table 1).

DV	Raw Alpha	Std Alpha
Free Will	0.81	0.81
Doxastic States	0.92	0.92
Conative States	0.93	0.93
Determinism	0.83	0.83
Moral Judgments	0.77	0.77

Table 1: Alpha Coefficients for the Five Pairs of Dependent Variables

As preregistered, we created five dependent variables (free will, doxastic states, conative states, moral judgment, determinism) by taking the means of the responses to the relevant pair of dependent variables. As was also preregistered, we analyzed the data by means of ANOVAs and conducted mediation analyses. We report the results in what follows.

4.3.1 Between-Subjects ANOVAs and Pairwise Comparisons

A series of between-subjects ANOVAs (see Appendix B for details[3]) determined, that aggregating across the Help and Hinder conditions participants were more inclined to ascribe free will ($F(1,778)= 550.88$, $p<.001$, $\eta p2=.41[.37,.46]$, a large effect), doxastic states ($F(1,778)= 563.72$, $p<.001$, $\eta p2=.42[.37,.46]$, a large effect), and conative states ($F(1,778)= 682.23$, $p<.001$, $\eta p2=.47[.42,.51]$, a large effect) to the marble in the Agent condition compared to the Object condition. Focusing on the Help Condition, participants further judged the marble morally better ($F(1,778)= 68.29$, $p<.001$, $\eta p2=.08[.05,.12]$, a medium effect) and to be less determined ($F(1,778)= 328.39$, $p<.001$, $\eta p2=.30[.25,.35]$, a large effect) in the Agent condition compared to the Object condition (Figure 5 for pairwise effects).

Moreover, participants were more inclined to ascribe free will ($F(1,778)= 8.43$, $p=.004$, $\eta p2=.01[.00,.03]$, a small effect) to the marble and judged the marble morally better ($F(1,778)= 41.16$, $p<.001$, $\eta p2=.05[.02,.08]$, a small effect), and to be more determined ($F(1,778)= 6.50$, $p=.011$, $\eta p2=.01[.00,.03]$, a large effect) in the Help condition than in the Hinder condition (Figure 6 for pairwise effects).

The results further revealed significant interactions for the attribution of free will ($F(1,778)= 53.82$, $p<.001$, $\eta p2=.06[.04,.10]$, a medium effect), doxastic mental states ($F(1,778)= 43.75$, $p<.001$, $\eta p2=.05[.03,.09]$, a small effect), and conative mental states ($F(1,778)= 33.24$, $p<.001$, $\eta p2=.04[.02,.07]$, a small effect) as well as on moral judgment ($F(1,778)= 28.23$, $p<.001$, $\eta p2=.04[.01,.06]$, a small effect).

[3] https://mfr.de-1.osf.io/render?url=https://osf.io/vn7kz/?direct%26mode=render%26action=download%26mode=render

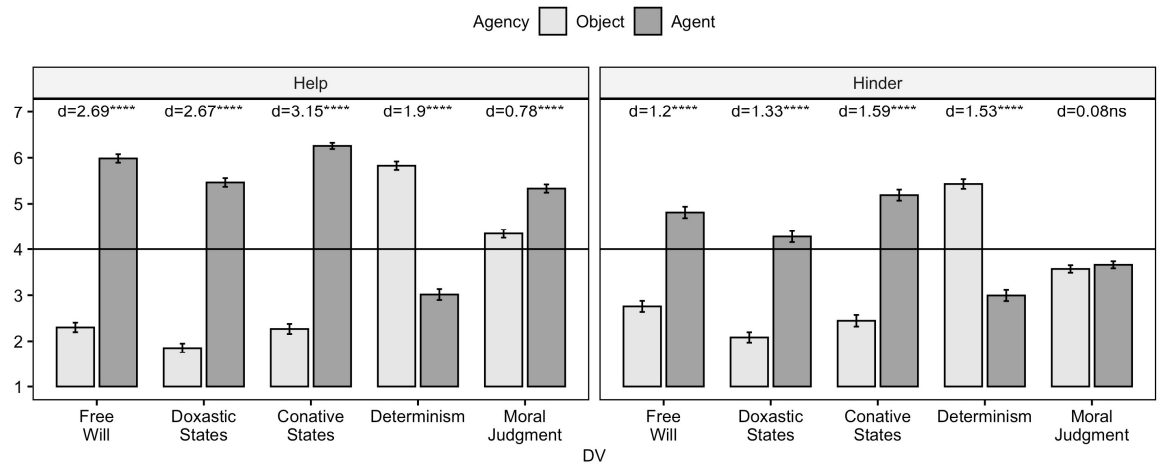

Figure 5: Mean Ratings for All Aggregated DVs across Agency (Agent vs. Object) and Valence (Help vs. Hinder). Error bars denote Standard Errors. Cohen's d for the Pairwise Effects of Agency (Agent vs. Object).

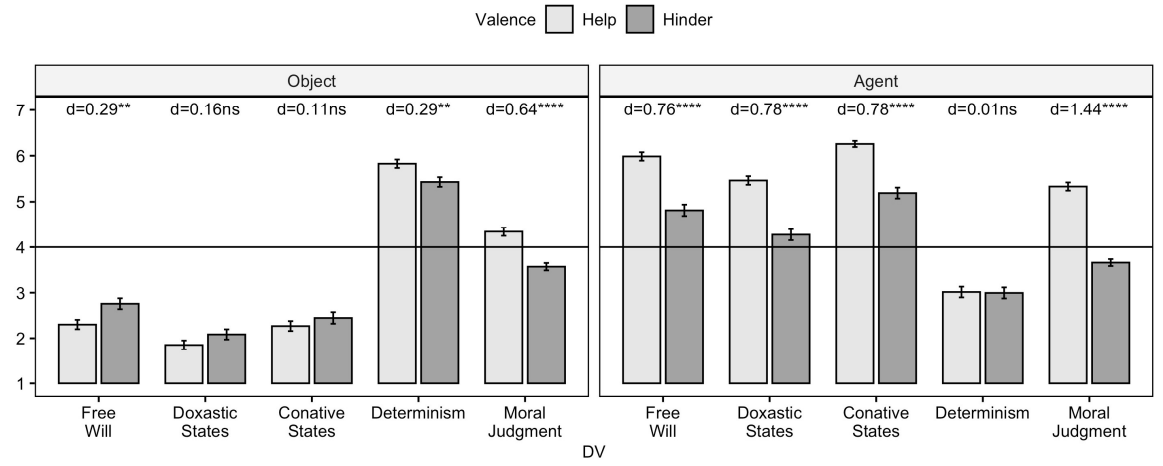

Figure 6: Mean ratings for all Aggregated DVs across Agency (Agent vs. Object) and Valence (Help vs. Hinder). Error bars denote Standard Errors. Cohen's d for the Pairwise Effects of Valence (Help vs. Hinder).

4.3.2. Mediation Analyses

To examine further the impact of Agency on free will, a multiple mediation analysis revealed Doxastic States and Conative States to be significant mediators (all *ps*<.009), whereas

Determinism and Moral Judgment proved nonsignificant (*p*s>.070) (Figure 7).[4] The impact of Agency on free will is largely mediated by the two significant factors, though a small, yet significant direct effect remained (*p*=.005).

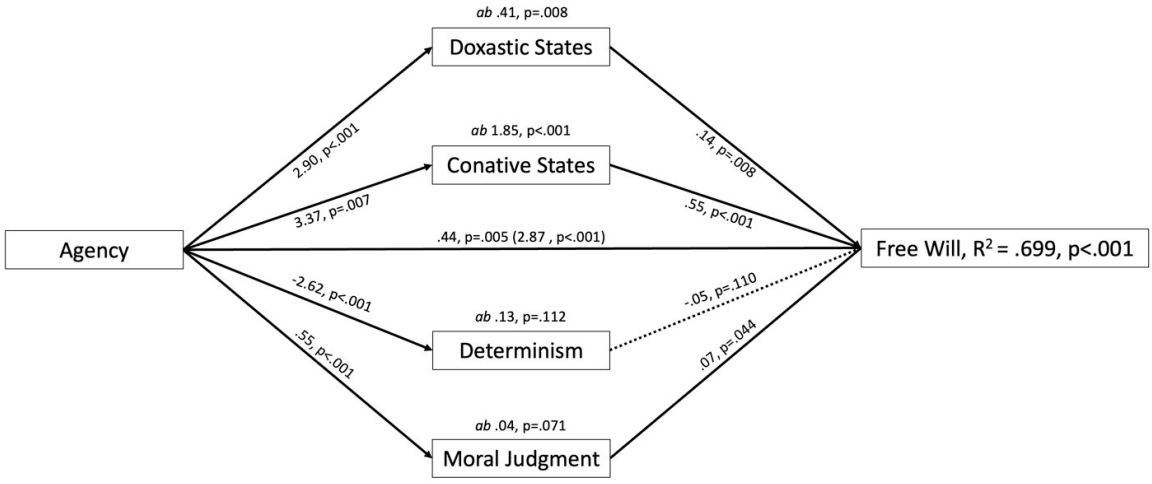

Figure 7: Mediation Analysis for the Impact of Agency on Free Will with Doxastic States, Conative States, Determinism, and Moral Judgment as Mediators.

[4] We had preregistered a simple mediation analysis with Determinism as a mediator. However, during data analysis we concldued that any effect observed in this analysis could be misleading if the other variables were not taken into account. In a mediation analysis with Determinism as the single potential mediator, the latter also proved nonsignificant as regards the relation between Valence and free will. It was significant for the relation between Agency and free will, although the indirect effect was rather small (see Appendix C: https://mfr.de-1.osf.io/render?url=https://osf.io/su5h9/?direct%26mode=render%26action=download%26mode=render). Further, the multiple mediation analysis suggests that Determinism is not a mediator.

As concerns the impact of Valence on free will, a multiple mediation analysis revealed a different pattern: All potential mediators proved significant (*p*s<.046) except for Determinism (*p*=.241) (Figure 8). Jointly, the three mediators render the significant impact of valence on free will nonsignificant (*p*=.618), suggesting that there is no direct effect of Valence on free will.

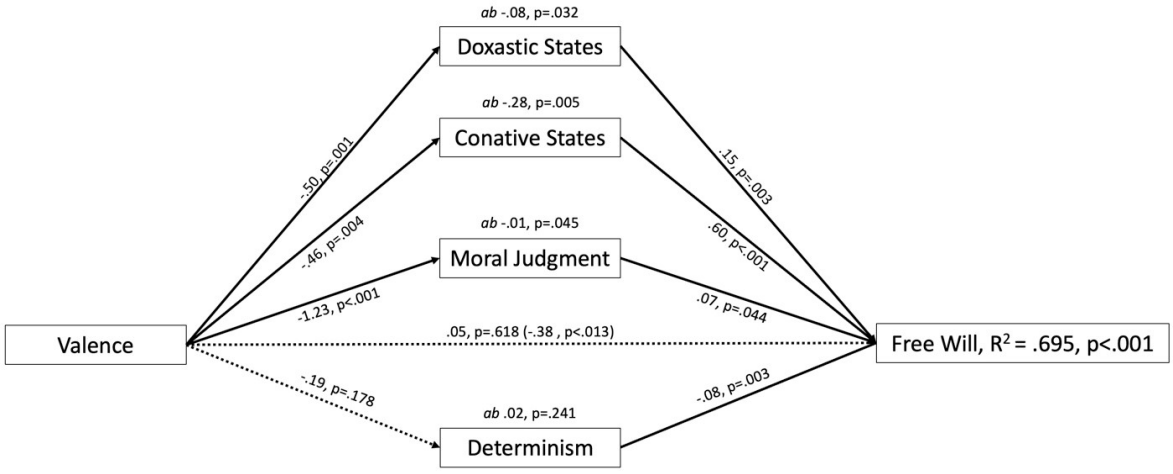

Figure 8: Mediation Analysis for the Impact of Valence on Free Will with Doxastic States, Conative States, Determinism, and Moral Judgment as Mediators.

Appendix F[5] reports the exploratory analyses specified in the preregistration (factor analysis and demographic analysis).

4.4 Discussion

In line with the agency model of free will attribution (prediction i), a simple geometrical figure, such as a marble, is viewed as minded and free when it embodies the agency cues that researchers

[5] https://mfr.de-1.osf.io/render?url=https://osf.io/u4dvr/?direct%26mode=render%26action=download%26mode=render

in the Heider and Simmel tradition have identified. Agency cues lead people to see objects as minded agents that are in control of their behavior. In line with the agency model (prediction ii), the presence of agency cues did not influence free will judgment by influencing judgments about whether the motion of the marble was physically and mechanistically determined. When it comes to the perception of free will, determinism does not seem to matter.

The valence of the marble's interaction with the mouse matters too, as revealed by the interaction between Valence and Agency for the attribution of free will, conative states, and doxastic states. Whether a mere object helps or hinders another agent in pursuing their goal makes no difference for the attribution of conative and doxastic states, as one would expect. By contrast, when an object behaves like an agent, people are more willing to assign free will, conative states, and doxastic states when it does something good, such as fostering someone's goals, than something bad, such as hindering someone's goals. This finding seems at odds with the large literature on the impact of moral valence on the attribution of conative and doxastic states as well as the application of other concepts, which has repeatedly found that people are more willing to assign doxastic and conative states when the agent does something bad. The results are also at odds with the smaller literature failing to find any effect of Valence when the stimuli are visual instead of verbal (Danks et al., 2014). We come back to this matter in the next section.

Finally, also noteworthy is the finding that the marble's behavior was judged to be morally similar in the Help and Hinder conditions when it is an agent. We also come back to this matter in the next section.

5. Perceiving Free Will on the Basis of Agency

5.1 Agency and Free Will

Our results provide support for the *agency model of free will attribution*. Adding agency cues, such as non-mechanical motion that is discontinuous and responsive, to a geometrical figure as simple

as a marble is sufficient to prompt people to not only see it as minded, that is, as having conative and doxastic states, but also as free. Not only do we reason to free will judgment, we also experience free will in our everyday life in relation to agency. This experimental finding is consistent with our own experiences of stimuli similar to Heider and Simmel's: We view the geometrical figures as being in control of their behavior.

These results illustrate the importance of going beyond the courtroom approach and of embracing the perceptual approach. While we undoubtedly reason about free will and control in some situations and while the features that determine whether we treat someone as free in such situations matter, in much of our everyday life free will attribution is a low-level process. We treat agents around us as free and in control either because agency, mindedness, free will, and perhaps even consciousness are attributed together in response to agency cues or because agency primes us to view agents as free.

What is more, the manipulation of agency cues did not influence free will judgments through judgments of determinism. Agency cues lead to free will judgment independently of reasoning about issues related to determinism. Much of the discussion of free will attribution in experimental philosophy has focused on whether determinism undermines free will attribution, and if so, why. Our results show that this focus is unfortunate since determinism does not influence our experience of free will.

5.2 The Psychology of Free Will

How do our results bear on the psychology of free will? Experimental philosophers have mostly been concerned with making explicit lay people's implicit theory of free will and, relatedly, responsibility (e.g., Hannikainen et al., 2019; Murray & Nahmias, 2014; Nahmias et al., 2005; Nichols, 2004; Nichols & Knobe, 2007; Sarkissian et al., 2010): Is this theory compatibilist? Why or why not? It is committed to the principle of alternate possibilties? On this view, people's understanding of free will has, just like contemporary philosophers', something to do with issues

related to determinism, although it is not settled whether laypeople are compatibilist, whether they are committed to the principle of alternate possibilities, whether the folk theory is universal or varies across cultures, etc. Experimental philosophers' approach to the lay concept of free will is echoed by others such as, e.g., Bloom (2012), who writes the following:

> [c]ommon sense tells us that we exist outside of the material world—we are connected to our bodies and our brains, but we are not ourselves material beings, and so we can act in ways that are exempt from physical law. (p. 1)

Like most experimental philosophers, Bloom connects the folk understanding of free will with the question of determinism (see also Wegner, 2002, 2003).

Clark, Baumeister, and colleagues reject this approach (e.g., Clark et al., 2014, 2017, 2019). On their view, lay people are committed to holding others responsible for their actions (in order to blame or, to a lesser extent, praise them) and as a result to treating them as free (for critical discussion, see Monroe & Ysidron, 2021). This view is inspired by Nietzsche, who wrote (1889/1954):

> Today we no longer have any pity for the concept of "free will": we know only too well what it really is—the foulest of all theologians' artifices, aimed at making mankind "responsible" in their sense. . . Wherever responsibilities are sought, it is usually the instinct of wanting to judge and punish which is at work. (p. 499)

Lay people embrace any view of free will that allows them to keep holding others free and responsible, even if it means embracing incompatible positions from one occasion to the next. As Clark and colleagues put it (2019), "people do not have one intuition about whether free will is compatible with determinism. Instead, people report that free will is compatible with determinism when desiring to uphold moral responsibility." That is, on this view, lay people do not really have a theory of free will, and experimental philosophers have mistakenly assumed they do.

Monroe and Malle (2010) have also rejected the experimental-philosophical approach to free will, which ties the folk conception of free will to issues related to determinism, but they hold, in contrast to Clark and colleagues, that there is a stable lay concept of free will that turns around the notion of choice and unconstrained action (see also; Feldman et al., 2014; Monroe et al., 2014, 2017; Stillman et al. 2011). Monroe and Malle (2010, p. 211) write that "the core of people's concept of free will is a choice that fulfills one's desires and is free from internal or external constraints. No evidence was found for metaphysical assumptions about dualism or indeterminism."

How do our results bear on these main psychological theories of free will judgment? First and foremost, none of these bodies of research embrace the perceptual approach that we have been touting in this chapter: They do not examine how perceptual cues can lead people to make free will judgment. Second, these bodies of research have little to say about the relation between agency and free will judgment. Third, the experimental-philosophical research and Clark and colleagues' theory both exaggerate the connection between the lay understanding of free will with decisions about how blame and praise should be apportioned. Similarly, as noted, the focus on the relation between determinism and free will attribution has misled many (thougb not all) researchers: Perceptual judgments about free will have nothing to do with determinism.

Our findings also seem to challenge Clark et al.'s (2014) asymmetry between praise and blame: On their view, free will attribution is particularly important to justify blame and punishment. What we found however is that more free will was assigned when the marble was viewed as an agent and helped compared to when it hindered someone's goal.

What does explain the suprising effect of praise and blame in our study? Stuart and Kneer (2021) have found that people assign more knowledge and blame to an autonomously acting robot when it doesn't commit any harm compared to when it commits harm. This "inverse outcome effect" mirrors the results found in our study. Stuart and Kneer propose that people are willing to assign mental states to robots and other entities that are not full-fledged

moral agents when there is little at stake in doing so; when the stakes are higher, for instance, when these entities would have to be blamed or punished, people are more reluctant to do so, treating them as the kind of things that can't be blamed or punished because they do not have mental states. This proposal can be easily combined with the agency model of free will attribution: People are primed to assign conative and doxastic states as well as free will in response to agency cues, but they can override or modulate this tendency when the stakes are higher (e.g., when blame is at stake). We speculate that stakes being higher is just one way of triggering reasoning-based judgments, which go beyond the more basic perception-based judgments that are the default in everyday life.

While we believe the challenge to Clark et al.'s theory is genuine and while we take the explanation just provided to be plausible, we should mention another possible explanation of why participants gave higher ratings to the free will, conative states, and doxastic states dependent variables in the Help compared to the Hinder condition. Despite our best efforts, the epistemic situation of the marble when it embodies agency cues is not the same in the Hinder and Help conditions. In the Help conditions the marble can see, so to speak, the mouse's efforts to open the door and it is pretty clear that the mouse wants to open the door. Someone watching the animated video can thus be fairly confident about what the marble knows and believes about the mouse as well as fairly confident about what it wants do to. In the Hinder condition, however, the marble doesn't have visual access to the mouse's efforts, and someone watching the animated video is probably less likely to be confident about the marble's conative and doxastic states. This epistemic asymmetry could explain the differences between the Help and Hinder conditions just discussed. Further research should balance the epistemic status of the marble better across conditions to find out whether this potential confound can explain away our findings.

5.3 Limitations

Our study is limited in some respects. We have already discussed a possible confound, but a second one should be mentioned too. On average participants gave similar answers to the moral judgment dependent variable in the Agent and Object conditions when the marble hinders the mouse's goal. Why is that? A possible explanation is that it isn't clear whether preventing a mouse from eating cheese is blameworthy and deserves punishment. If it isn't clear, people should give an average answer near the indifferent point, which they should also do for the object condition. This hypothesis could explain why we failed to find any difference in moral judgment between the Agency and Object conditions when the marble hinders' the mouse's goal. In addition, the study didn't examine how the agency cues trigger a free will attribution nor did it examine in what sense people perceive or experience free will. More research is called for about these issues. Finally, we do not deny that reasoning plays a role in free will judgment. Reasoning and low-level cues might interact in a complex manner even in the kind of situations used in this study. Future work should investigate this interaction.

Conclusion

This chapter has made a plea for a new approach to free will attribution: the perceptual approach. We not only reason to free will, the topic of most of the experimental philosophy of free will, but we also see free will. We have proposed that the cues that lead people to see agency also lead them to see free will. Visual stimuli were developed to study the attribution of free will that is grounded in perception, and our results show that agency cues matter for free will attribution. Finally, these results reveal that concerns with determinism and with moral judgment that have been central to much of the psychology of free will judgment are not central to our everyday experience of free will.

Acknowledgements

Pascale Willemsen's research was supported by the Swiss National Science Foundation, Grant Number: PCEFP1_181082.

Markus Kneer's research was supported by the Swiss National Science Foundation, Grant Number: PZ00P1_179912.

Albert Newen's research was supported by Deutsche Forschungsgemeinschaft (DFG) – Projekt number GRK-2185/1 (DFG-Graduiertenkolleg Situated Cognition) and by the DFG-project (NE 576/14-1) "The structure and development of understanding actions and reasons".

References

Arico, A, Fiala, B., Goldberg, R. F., & Nichols, S. (2011). The Folk Psychology of Consciousness. *Mind & Language, 26*(3), 327–352. https://doi.org/10.1111/j.1468-0017.2011.01420.x

Baillargeon, R., Scott, R. M. & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14(*3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Barrett, 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. https://doi.org/10.1016/j.evolhumbehav.2004.08.015

Beebe, J. & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, *25(*4), 474–498. https://doi.org/10.1111/j.1468-0017.2010.01398.x

Bloom, P. (2012, March 18). Free will does not exist. So what? *The Chronicle Of Higher Education.* https://www.chronicle.com/article/free-will-does-not-exist-so-what

Bowler, D. & Thommen, E. (2000). Attribution of Mechanical and Social Causality to Animated Displays by Children with Autism. *Autism*, *4*(2), 147–171. https://doi.org/10.1177/1362361300004002004

Carruthers, P. (2015). Perceiving mental states. *Consciousness and Cognition*: *An International Journal, 36*, 498–507. https://doi.org/10.1016/j.concog.2015.04.009/

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: a motivated account of free will belief. *Journal of Personality and Social Psychology, 106*(4), 501–513. https://doi.org/10.1037/a0035880

Clark, C. J., Baumeister, R. F., & Ditto, P. H. (2017). Making punishment palatable: Belief in free will alleviates punitive distress. *Consciousness and Cognition*: *An International Journal*, *51*, 193–211. https://doi.org/10.1016/j.concog.2017.03.010

Clark, C. J., Winegard, B. M., & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology*, *10*, Article 215. https://doi.org/10.3389/fpsyg.2019.00215

Csibra, G., Gergely, G., Bíró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, *72*(3), 237–267. https://doi.org/10.1016/S0010-0277(99)00039-6

Danks, D., Rose, D. & Machery, E. (2014). Demoralizing causation. *Philosophical Studies*, *171*(2), 251–277. https://doi.org/10.1007/s11098-013-0266-8

Feldman, G., Baumeister, R. F. & Wong, K. F. E. (2014). Free will is about choosing: The link between choice and the belief in free will. *Journal of Experimental Social Psychology*, *55*, 239–245. https://doi.org/10.1016/j.jesp.2014.07.012

Feltz, A., & Cova, F. (2014). Moral responsibility and free will: A meta-analysis. *Consciousness and Cognition*: *An International Journal, 30*, 234–246. https://doi.org/10.1016/j.concog.2014.08.012

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*: *An International Journal, 17*(2), 535–543. https://doi.org/10.1016/j.concog.2008.03.003

Gallese, V. & Goldman. A. I. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501. https://doi.org/10.1016/S1364-6613(98)01262-5

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The Wolfpack Effect: Perception of Animacy Irresistibly Influences Interactive Behavior. *Psychological Science*, *21*(12), 1845–1853. https://doi.org/10.1177/0956797610388814

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607. https://doi.org/10.1037/xge0000670

Goldman, A. I. (1992). In defense of the simulation theory. *Mind & Language*, *7*(1-2), 104–119. https://doi.org/10.1111/j.1468-0017.1992.tb00200.x

Goldman, A. I. (2006). *Simulating Minds. The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford University Press. https://doi.org/10.1093/0195138929.001.0001

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language, 7*(1-2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Hannikainen, I. R., Machery, E., Rose, D., Stich, S., Olivola, C. Y., Sousa, P., Cova, F., Buchtel, E., Alai, M., Angelucci, A., Berniûnas, R., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg … Zhu, J. (2019). For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology*, *10*, 2428. https://doi.org/10.3389/fpsyg.2019.02428

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

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *The Journal of Philosophy*, *106*(11), 587–612. https://doi.org/10.5840/jphil20091061128

Hollander-Blumoff, R. E. (2012). Crime, Punishment, and the Psychology of Self-Control. *Emory Law Journal, 61*(3), 501–552.

Kneer, M. (2018). Perspective and Epistemic State Ascriptions. *Review of Philosophy and Psychology*, *9*(2), 313–341. https://doi.org/10.1007/s13164-017-0361-4

Kneer, M. (2021). Reasonableness on the Clapham Omnibus: Exploring the folk concept of reasonable. In Bystranowski, P., Janik, B. & Prochnicki, M. (Eds.), *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives*. Springer Nature.

Kneer, M., Colaço, D., Alexander, J., & Machery, E. (in press). On second thought: Reflection on the reflection defense. In T. Lombrozo, J. Knobe & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy, Volume 4*. Oxford University Press.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, *63*(3), 190–194. https://doi.org/10.1111/1467-8284.00419

Monroe, A. E. & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, *1*, 211–224. https://doi.org/10.1007/s13164-009-0010-7

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing Free Will down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment. *Consciousness and Cognition*: *An International Journal, 27*, 100–108. https://doi.org/10.1016/j.concog.2014.04.011

Monroe, A. E., Brady, G. L., & Malle, B. F. (2017). This Isn't the Free Will Worth Looking For: General Free Will Beliefs Do Not Influence Moral Judgments, Agent-Specific Choice Ascriptions Do. *Social Psychological and Personality Science*, *8*(2), 191–199. https://doi.org/10.1177/1948550616667616

Monroe, A. E., & Ysidron, D. W. (2021). Not so motivated after all? Three replication attempts and a theoretical challenge to a morally motivated belief in free will. *Journal of Experimental Psychology: General*, *150*(1), e1–e12. https://doi.org/10.1037/xge0000788

Murray, D., & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434–467. https://doi.org/10.1111/j.1933-1592.2012.00609.x

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying Freedom: Folk Intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584. https://doi.org/10.1080/09515080500264180

Newen, A., Welpinghus, A., & Juckel, G. (2015). Emotion recognition as pattern recognition: the relevance of perception. *Mind & Language*, *30*(2), 187–208. https://doi.org/10.1111/mila.12077

Newman, G. E., Bloom, P. & Knobe, J. (2014). Value Judgments and the True Self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Nichols, S. & Stich, S. (2003). *Mindreading*: *An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press. https://doi.org/10.1093/0198236107.001.0001

Nichols, S. (2004). The Folk Psychology of Free Will: Fits and Starts. *Mind & Language*, *19*(5), 473–502. https://doi.org/10.1111/j.0268-1064.2004.00269.x

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs*, *41*(4), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

Nietzsche, F. (1954). *Twilight of the idols* (W. Kaufmann, Trans.). Penguin Books. (Original work published 1889)

Reid, V. M., Csibra, G., Belsky, J., & Johnson, M. H. (2007). Neural correlates of the perception of goal-directed action in infants. *Acta psychologica*, *124*(1), 129–138. https://doi.org/10.1016/j.actpsy.2006.09.010

Rose, D., Buckwalter, W., & Nichols, S. (2017). Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics. Cognitive Science, 41(2), 482–502. https://doi.org/10.1111/cogs.12310

Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (2020). Natural Compatibilism, Indeterminism, and Intrusive Metaphysics. *Cognitive Science*, *44*(8), Article e12873. https://doi.org/10.1111/cogs.12873

Santos, N. S., David, N., Bente, G., & Vogeley, K. (2008). Parametric induction of animacy experience. *Consciousness and Cognition*: *An International Journal, 17*(2), 425–437. https://doi.org/10.1016/j.concog.2008.03.012

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is Belief in Free Will a Cultural Universal? *Mind & Language*, *25*(3), 346–358. https://doi.org/10.1111/j.1468-0017.2010.01393.x

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment? In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 197–230). MIT Press. https://doi.org/10.7551/mitpress/9780262019279.001.0001

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Stillman, T. F., Baumeister, R. F., & Mele, A. R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology, 24*(3), 381–394. https://doi.org/10.1080/09515089.2011.556607

Stuart, M. T. & Kneer, M. (2021). Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. *Proceedings of the ACM on Human-Computer Interaction, 5*, Article 363. https://doi.org/10.1145/3479507

Sytsma, J. (2020). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, 1–21. https://doi.org/10.1007/s13164-020-00498-2

Tannenbaum, D., Ditto, P. H., & Pizarro, D. A. (2007). *Different moral values produce different judgments of intentional action* [Unpublished manuscript]. University of California-Irvine. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.306.9800&rep=rep1&type=pdf

Wegner, D. M. (2002). *The Illusion of Conscious Will*. MIT Press.

Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences, 7*(2), 65–69. https://doi.org/10.1016/S1364-6613(03)00002-0

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass, 14*(1), Article e12562. https://doi.org/10.1111/phc3.12562