Force Dynamics as a Basis for Moral Intuitions

Jonas Nagel (jnagel1@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

People seamlessly generate moral intuitions about a wide range of events they observe, but to date the cognitive processes underlying this competency are poorly understood. We propose that our moral intuitions are grounded in force-dynamic intuitions. We show how the evaluation of entities engaged in schematized interactions can be predicted from specific force-dynamic properties of those interactions, and we point out how these evaluative tendencies relate to our moral norm of not interfering with others' interests.

Keywords: moral judgment; intuition; force dynamics

A New Theory of Moral Intuitions

Recent moral psychology views intuitions as important determinants of our moral judgments. Haidt (2001) defined moral intuitions as "the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion" (p. 818). Intuitions are thus mainly defined in contrast to deliberative reasoning.

However, to date there is no worked-out theory of the automatic processes by which our moral intuitions are formed. How do we solve this computational task? Which observed events elicit which specific moral intuitions? In what format are these events represented, and how is this representation automatically integrated with pre-existing evaluative standards? We propose that the semantic category of *force dynamics* (Talmy, 1988) provides a cognitive structure that might serve as a representational basis in this task.

The Semantic Category of Force Dynamics

Talmy (1988) described force dynamics as a semantic category of how entities interact with respect to force. When two entities interact in the world, our language assigns to them the two thematic roles of *patient* (*P*) and *agent* (*A*). (Talmy uses the terms *agonist* and *antagonist*.) *P*, the focal entity, is perceived to have an intrinsic force tendency either towards rest or towards motion. In force-dynamic interactions, *P* finds itself in opposition to *A*, another entity with the opposed force tendency. *A* is mainly thought of in terms of the effects it has on the resultant force manifested by *P* as outcome of the interaction.

P's resultant force depends on the relative strength of the two opposing forces. If *P*'s force is stronger than *A*'s, then *P*'s resultant force equals its intrinsic tendency. This constellation is expressed in familiar words in our natural language, such as *despite* or *although*. In "The flame [*P*] kept burning despite the wind [*A*] blowing at it," *P* manifests its intrinsic tendency (to burn) in spite of the opposing force exerted by *A*.

If *A*'s force is stronger than *P*'s, then *P*'s resultant force is opposed to its intrinsic tendency. There are many words in natural language describing variants of this basic constellation (e.g., *make*, *cause*, or *prevent*). In the sentence "The wind [*A*] made the flame [*P*] go out," *P* does not manifest its intrinsic tendency (to burn) but the opposite (to go out). Note that both example sentences "are about" *P*, while *A* is mainly relevant in terms of the effect it has on *P*.

Talmy (1988) argues that force dynamics is a fundamental semantic category, profoundly structuring our cognitions in a variety of domains. Whether we deal with the physical, the (intra-)psychic, or the social world, the same basic force-dynamic concepts pervade our language and thought. Thus, force dynamics is conceived as a domain-independent representation underlying intuitions in various domains, not only in the physical domain. Actually it is interesting to see that when force dynamics is applied to physical tasks, the resulting intuitions about forces and intrinsic tendencies seem to be more compatible with our understanding of actions than with Newtonian physics. For example, the intuition that causes have stronger forces than effects is inconsistent with Newtonian physics but seems to be grounded in sensory-motor representations of our actions (White, 2009). Analyzing social interactions in terms of force dynamics is thus not a reduction of the social to the physical domain. Force dynamics is better viewed as a domain-independent *abstract* conceptual framework.

Force Dynamics as a Basis for Moral Intuitions

The category of force dynamics combines causal and teleological aspects and is abstract enough to be naturally applicable across physical and social domains. These properties make it a promising candidate to serve in the process of enriching representations of observed events with a basic evaluative aspect.

Imagine observing the following event: Jack shoves Jones. In a first step, the observer could abstract the force-dynamic pattern instantiated by this event. This would include assigning *A*- and *P*-roles to the entities involved in the interaction, determining *P*'s intrinsic tendency and resultant force, and comparing the latter two. In this example, this would yield a representation of *A* (Jack) forcing *P* (Jones) to deviate from his intrinsic tendency (toward rest) into a different resultant force (motion). This is an instance of *onset causing of motion* (Talmy, 1988).

In a second step, this abstract representation could be automatically subjected to default normative principles formulated on the same level of abstraction. We stipulate the existence of a *noninterference principle (NIP)*: By default, *patients should be allowed to manifest their intrinsic tendencies*. This substantive assumption (which has itself a force-dynamic structure) can be motivated with reference to the negative prima facie duty not to interfere with others' interests, which seems to be a fundamental moral norm at least in Western cultures. In *onset causing of motion*, this abstract principle is violated. *P* changes its tendency because of *A*'s impingement.

Finally, the fact that *A* is identified as causing a violation of the principle leads people to evaluate *A* negatively relative to *P*. This negative evaluation is then applied to the concrete observed instance of *A* (in this case, to Jack).

In the current research we focus on basic scenarios with only two protagonists (*A*, *P*). Obviously, there are many more complex instances of *onset causing* patterns in which *A* might eventually be evaluated positively. Imagine you receive the additional information that Jack shoved Jones *out of harm's way*. Such more complex constellations will not be treated here, but it seems that our theory can in principle be extended to handle them as well (e.g., Jack could be evaluated positively for *preventing* another agent [the harm] to violate the NIP by means of intervening on *P*).

In what follows, we will provide initial evidence that such default evaluations are in fact assigned on the abstract level of decontextualized force-dynamic representations. To this end, we had experimental subjects evaluate the movements of two abstract shapes engaged in simple interactions.

Force Dynamics in Abstract Animated Displays

Displays of moving objects allow for a non-verbal presentation of decontextualized force-dynamic interactions. In the absence of linguistic cues, we first need to explicate the criteria according to which we assume our subjects to abstract force-dynamic patterns from our visual displays (i.e., the first step in the process outlined above).

We instantiated the *onset causing of motion* pattern with a version of the well-known *launching event* (Michotte, 1963; see Fig. 1). A stationary Object Y is situated in the center of the stage. After a moment, another Object X enters the scene from the side and approaches Object Y on a straight line and at constant speed. At the moment of contact, Object X stops and Object Y immediately starts moving as if it was continuing on X's trajectory.

According to our theory, subjects first need to assign agent and patient roles to these interacting entities. We argue that the extremely impoverished nature of this display leaves but three cues to make this assignment: (a) pre-collision movement relative to the position of the other entity; (b) causing change of state in the other entity; and (c) appearing on the scene after the other entity. According to Dowty (1991), the first two cues increase the likelihood that a given entity is assigned the agent role in an interaction. Concerning the third cue, the entity appearing first will likely be seen as the focal entity the display "is about" (i.e., the patient); the second entity should therefore be regarded as agent affecting this focal entity. We assign the agent role to an entity if it embodies more of these three cues than the alternative entity. In the launching event, Object X's behavior is consistent with all three cues (a+, b+, c+), while Object Y only causes Object X to stop (b+), but does not move initially (a-) and is first on the screen (c-). Object X is therefore assigned the agent role, while Object Y is the patient. There is ample evidence that this analysis is in line with people's qualitative experience of launching events. For example, X is seen as *exerting force on* Y, whereas Y is perceived to merely *exhibit resistance against* X (White, 2009).

Next, subjects need to infer *P*'s intrinsic tendency and resultant force. We assume that in the absence of further contextual cues indicating the presence of external forces, *P*'s pre-collision movement will be regarded as its intrinsic tendency. The identification of *P*'s resultant force with its post-collision movement seems unproblematic.

Finally, *P*'s intrinsic tendency and resultant force need to be compared in order to determine the force-dynamic pattern and to decide whether the NIP was violated, as would be indicated by a change of *P*'s movement as a result of the collision event.

Hypothesis

With all force-dynamic concepts operationalized, we now turn to the specifics of the hypothesis we tested in the present experiment. We predict that in the *Launch* case described above, which instantiates *onset causing of motion*, subjects asked to evaluate the movements of both entities on a negative/positive dimension will evaluate Object X more negatively than Object Y. We contrast this case with a *Blocked* case which is identical to *Launch* except that Object Y does not start moving on X's trajectory after the collision, so that the interaction ends with both entities at rest in the center of the screen. In this case, X has two agentic cues (a+, b-, c+) while Y has only one (a-, b+, c-). Object Y is thus still the patient, and its intrinsic tendency (rest) is identical to its resultant force (rest). This corresponds to a *despite* pattern in Talmy's (1988) terminology. The NIP is not violated here, so we do *not* expect *A* (Object X) to be rated negatively relative to *P* (Object Y) in this case. Across the cases, we expect *A* to be rated more negatively in *Launch* than in *Blocked*.

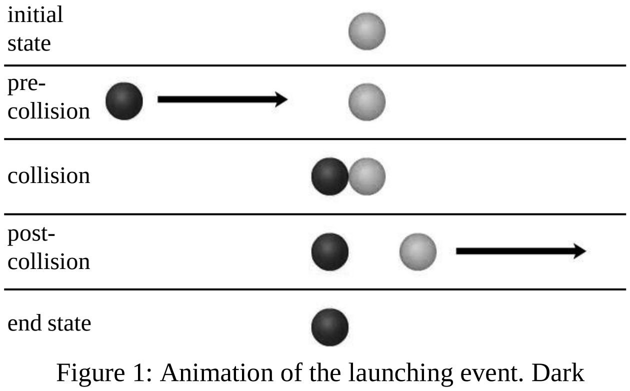

Figure 1: Animation of the launching event. Dark sphere = *A*; light sphere = *P*. See text for details.

Experiment

We divide the presentation of the experiment into two parts. In the first part, we test the hypothesis just explicated. The second part will deal with an additional aspect. However, the data for both parts were gathered within one and the same experiment and from the same subjects. Therefore, we begin by describing the general procedure for the whole experiment before we discuss the specific materials and the results for both parts separately.

Participants and Procedure

31 undergraduates of the University of Göttingen (23 female, mean age 22 years) participated in a computer-based experiment containing 27 trials in random order.[1] Each trial consisted of three consecutive screens. The first screen displayed the trial number and a button with which the subjects could start the animation. The second screen contained a looped display of one of 27 animations. Each animation started with a fixation cross displayed for one second in the center of the stage. Then the initial state was presented for one second, before the force dynamic interaction began to unfold. The interaction always consisted of a blue sphere and a green sphere moving in specific ways on the same straight horizontal trajectory. After the interaction was finished, the stationary end state remained on the screen for one second. As an example, Figure 1 illustrates how we implemented the *Launch* case described above. Each condition was instantiated in two animations, counterbalanced for color assignment and direction of movement. Subjects watched each animation as often as they wished before they proceeded to the third screen, where two identical 7-point rating scales ranging from "-3 (negative)" to "+3 (positive)" were presented above one another. Each referred to one of the two entities from the animation. The question wording was "How do you evaluate the movement of the blue/green figure?"

Part 1: Interference with Intrinsic Tendency

Design and Material In this part we tested whether people's intuitive evaluations of the entities in *Launch* and *Blocked* can be predicted from the underlying force dynamics in combination with the noninterference principle (NIP). In the initial state of all animations, *P* was displayed at rest in the center of the stage. The force-dynamic interaction always began with *A* entering from one side and reaching *P* within one second on a straight horizontal trajectory at constant speed. We then manipulated *A*'s and *P*'s post-collision movement which could either be stationary (0), movement continuing *A*'s initial trajectory at half of *A*'s initial speed (1), or movement on the same trajectory at *A*'s full initial speed (2). Both *A* and *P* could display all three movements, with the constraint that *A* could not be faster than *P* after collision because this would imply the objects going through each other. Thus, the manipulation yielded six conditions which are displayed in Table 1. Conditions 1 and 3 are of main theoretical interest because they manifest the *Blocked* and *Launch* cases to which our main hypothesis refers.

Table 1 also lists several resulting properties of the interactions displayed in each condition. *A*_change and *P*_change indicate whether *A* and *P* change their overt tendency in the course of the interaction. *P*_change is the most important variable for our purpose. Given the pre-collision constellation (i.e., *A* in motion, *P* at rest), as soon as *P* changes its tendency, the case is an instance of *onset causing of motion*, and the NIP is violated. Only if *P* stays stationary, the case becomes an instance of *despite* where the principle is not violated. The remaining properties are further implications of the entities' post-collision movements. Concordance indicates whether *A* and *P* have a concordant tendency (either of rest or of movement in the same direction) after the collision. Contact indicates whether *A* and *P* remain in direct physical contact after the collision. Finally, Resistance indicates whether *P* displays resistance by not overtaking the pre-collision tendency of *A* in a one-to-one manner. As can be seen in Table 1, this property is not identical to *P*_change.

Concordance, Contact, and Resistance are listed because they are perfectly confounded with *P*_change across the two cases of main interest, 1 and 3. Any change in ratings between 1 and 3 could thus just as well be caused by these properties. The remaining four conditions serve to isolate *P*_change from these confounds.

Table 1: Design of Part 1

	Speed		Resulting properties				
Cond	*A*	*P*	*A*_ch	***P*_ch**	Conc	Cont	Res
1	**0**	**0**	**1**	**0**	**1**	**1**	**1**
2	0	1	1	**1**	0	0	1
3	**0**	**2**	**1**	**1**	**0**	**0**	**0**
4	1	1	1	**1**	1	1	1
5	1	2	1	**1**	1	0	0
6	2	2	0	**1**	1	1	0

Note. Cond = condition, Speed = post-collision speed, *A*/*P*_ch = *A*/*P*_change, Conc = Concordance, Cont = Contact, Res = Resistance. See text for further explanations.

Specific predictions Three specific predictions follow directly from our hypothesis. (i) *A* will be rated more negatively than *P* in condition 3 (*Launch*). (ii) *A* will *not* be rated more negatively than *P* in condition 1 (*Blocked*). (iii) *A* will be rated more negatively in 3 than in 1.

The remaining conditions serve to separate the manipulation of the force-dynamic pattern from the properties of Concordance, Contact, and Resistance. If a concordant post-collision tendency is responsible for more positive ratings for *A* in 1 compared to 3, *A*-ratings should also be more positive in 4, 5, and 6. If sustained physical contact is to be made responsible, *A*-ratings should be more positive in 4 and 6. Finally, if the display of resistance by *P* (in not adopting *A*'s pre-collision tendency) is to be made

[1] Three of these 27 trials tested a third hypothesis which is not reported here due to space constraints.

responsible, *A*-ratings should be more positive in 2 and 4. We predict that none of these alternatives will be the case. Instead, we expect (iv) all control cases to be treated like 3 since they all conform to the *onset causing of motion* pattern. This result would support our hypothesis that the evaluative ratings in 1 (*Blocked*) and 3 (*Launch*) are in fact a function of the underlying force-dynamic pattern as indicated by the variable *P*_change.

Results and Discussion The descriptive results are displayed in Figure 2. A global 6 (Condition: 1 to 6) × 2 (Entity: *A* vs. *P*) repeated-measures ANOVA yielded a main effect for Entity ($F_{1,30} = 9.57$; $p < .01$, $\eta_p^2 = .24$), indicating that, across conditions, *A* was rated more negatively than *P*. More importantly, the Condition × Entity interaction term was significant ($F_{5,150} = 9.69$; $p < .001$, $\eta_p^2 = .24$), showing that *A* and *P* were treated differently across conditions. We now turn to the contrasts testing our specific predictions.

(i): In 3 (*Launch*), *A*-ratings were lower than the *P*-ratings ($t_{30} = -3.68$, $p < .001$, $d = -.66$). *A* is thus rated more negatively than *P* in the launching event as paradigmatic instance of the *onset causing of motion* pattern violating the NIP.

(ii): In 1 (*Blocked*), *A*-ratings were *higher* than *P*-ratings ($t_{30} = 3.63$, $p < .01$, $d = .65$). Thus, it seems that *A*'s negative evaluation disappears with an underlying *despite* pattern in which the NIP is not violated. The significant drop in *P*-ratings was not expected because our predictions only concerned *A*-ratings. One post-hoc explanation for this phenomenon might be that the agent-patient distinction is not as clear cut in this case as in the *Launch* case (i.e., the agentic cues are distributed more evenly across both entities). Maybe some participants interpreted *P* as agent due to its capacity to cause change in *A*, turning the interaction into an *onset causing of rest* pattern (Talmy, 1988) in which *P* (now the agent) violates the NIP by forcing *A* (now the patient) to deviate from its intrinsic tendency to motion into a resultant state of rest.

(iii): *A*-ratings in 1 were higher than those in 3 ($t_{30} = 3.27$, $p < .01$, $d = .59$). Our hypothesis is thus confirmed by comparisons between entities within cases and across cases with different underlying force-dynamic patterns.

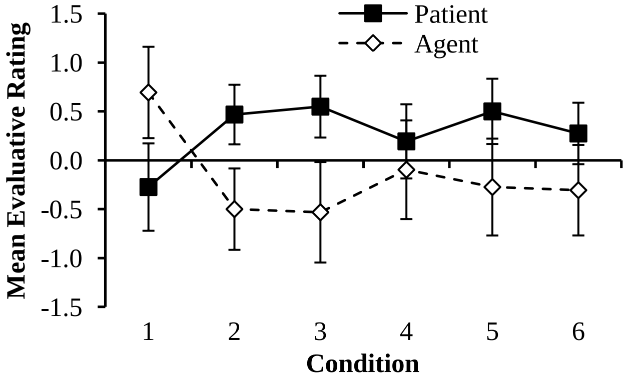

Figure 2: Results of Part 1. Error Bars = 95% CI.

(iv): The significant increase in *A*-ratings only occurred in 1, while, as predicted, the four control conditions generally behaved like 3 (*Launch*). The only exception is condition 4, where the *A*-ratings were not significantly lower than the *P*-ratings (although the descriptive trend still holds, $t_{30} = -1.31$). Furthermore, across all six conditions, there seems to be a trend for concordant post-collision movement (i.e., 1, 4, 5, and 6) to yield slightly higher *A*-ratings than discordant post-collision movement ($t_{30} = 2.26$, $p < .05$, $d = .41$). This might indicate that Concordance is used as an additional cue for the evaluation of *A*. However, note that this effect is driven mainly by the selective increase of *A*-ratings in 1 (*Blocked*).

In sum, these findings demonstrate that in clear-cut cases of *onset causing of motion* such as 3 (*Launch*), agents are evaluated more negatively than patients. This is not the case in *despite* cases in which the NIP is not violated. Together, these findings show that people's evaluations of movements are sensitive to underlying force-dynamic patterns. Entities are by default evaluated negatively if they cause other entities to deviate from their intrinsic tendency. This result is consistent with the moral norm in our society to not force others into states in which they would not enter on their own.

Part 2: Prior Concordance

In this part we will test whether the default evaluations we have discovered are robust enough to be consistently applied to cases in which *P* is initially not at rest but rather in motion. Imagine a moving *P* colliding with a faster-moving *A*, changing the speed and/or direction of its movement after the collision. According to our criteria, *P*'s intrinsic tendency would be to move in exactly the manner manifested prior to collision (including direction and speed parameters). *A* should thus be identified as causing *P* to deviate from its intrinsic tendency which constitutes a violation of the noninterference principle (NIP).

Crucially, this should be the case regardless of whether *P* and *A* exhibit concordant or discordant pre-collision movement. The direction of forces is not represented in Talmy's (1988) framework so that two entities moving on the same trajectory at different speed are still conceptualized as being in opposition (contrary to Wolff, 2007; see General Discussion). Note that without reference to Talmy's framework of opposing forces it seems a priori plausible that the concordant and discordant cases will be conceptualized differently. Specifically, in the concordant case it may seem as if the faster *A* *enhances* the slower *P* in its tendency which could result in *A* being evaluated positively for "helping" *P*. According to our theory, however, this should not be the case. If people's default evaluations correspond to Talmy's framework and the NIP, then they should be insensitive to prior concordance: They should evaluate an *A* making *P* go faster into the direction of its initial movement similarly to an *A* making *P* go into the opposite direction of its initial movement. Both cases violate the NIP.

Design and Material The initial state of all animations was an empty stage. After one second, *P* entered the stage from one side at a constant speed on a straight horizontal trajectory, reaching the center of the stage after two seconds. One second after *P*'s appearance, *A* entered the stage at twice the speed of *P*, either from the same side (concordant condition, C+) or from the opposite side (discordant condition, C-). Consequently, in both conditions the collision of *A* and *P* took place in the center of the screen, one second after *A*'s appearance. After the collision, both entities moved in the direction of *A*'s initial movement in all conditions. This implies that in C+, *P* continued in the direction of its initial movement, while in C- it reversed the direction of movement. Similar as in Part 1, we manipulated the post-collision speed of both entities, which could be the initial speed of *P* (1) or *A* (2). Again, *P* had to be at least as fast as *A* after the collision. This yielded three Speed conditions crossed with the two Concordance conditions, resulting in the six experimental conditions summarized in Table 2. Concerning the definition of agentic cues, we refined the criterion of pre-collision movement (a, see above) to *faster* pre-collision movement relative to the other entity. As in Part 1, *A* exhibits at least one more cue for agency than *P* in all conditions.

Table 2: Design of Part 2

		Speed		Resulting Properties			
Cond	**Conc**	*A*	*P*	*A*_ch	***P*_ch**	Cont	Res
C+1	**1**	1	1	1	**0**	1	1
C+2	**1**	1	2	1	**1**	0	0
C+3	**1**	2	2	0	**1**	1	0
C-1	**0**	1	1	1	**1**	1	1
C-2	**0**	1	2	1	**1**	0	0
C-3	**0**	2	2	0	**1**	1	0

Note. Cond = Condition, Conc = pre-collision Concordance, Speed = post-collision speed, *A*/*P*_ch = *A*/*P*_change, Cont = Contact, Res = Resistance.

Specific Predictions (i) Straightforward predictions arise from our model for the evaluation of *A* in all three C- cases. The reversal of *P*'s direction of movement caused by *A* is a clear violation of the NIP which should lead subjects to evaluate *A* negatively relative to *P*.

(ii) Case C+1 is an instantiation of *despite* in which *P* continues manifesting its intrinsic tendency after the collision. *A* should not be evaluated negatively relative to *P* since the NIP is not violated.

(iii) The crucial new conditions are C+2 and C+3. Here, subjects encounter a violation of the NIP preceded by concordant pre-collision movement. *P* is thus merely caused to deviate from its intrinsic (slow) speed, but not to deviate from its intrinsic direction. This could in principle lead subjects to conceptualize *A* as helping *P* to advance faster on its path. However, our theory predicts that subjects will still evaluate *A* negatively for causing a violation of the NIP. The *A*-ratings should also not differ from the *A*-ratings in the C- conditions.

Results and Discussion The descriptive results are displayed in Figure 3. A global 2 (Concordance: C+ vs. C-) × 3 (Speed: 1 to 3) × 2 (Entity: *A* vs. *P*) repeated-measures ANOVA yielded a main effect for Concordance ($F_{1,30} = 19.10$; $p < .001$, $\eta_p^2 = .39$), indicating that both entities were generally rated more negatively in C- than in C+. Again there was a main effect for Entity ($F_{1,30} = 28.90$; $p < .001$, $\eta_p^2 = .49$), indicating that *A* was generally rated more negatively than *P*. Finally, the Speed × Entity and the Speed × Concordance interaction terms were significant ($F_{1,30} = 7.08$; $p < .01$, $\eta_p^2 = .19$, and $F_{1,30} = 4.07$; $p < .05$, $\eta_p^2 = .12$, respectively), showing that post-collision movements of *A* and *P* affected *A*- and *P*-ratings differentially, and that they also had different effects depending on the prior concordance of both entities.

(i) *A*-ratings in the three C- conditions are lower than the respective *P*-ratings ($t_{30} = -4.02$, $p < .001$, $d = -.72$), replicating the result of Part 1 that *A* receives negative evaluations when it clearly violates the NIP.

(ii) *A*-ratings in C+1 are not different from the *P*-ratings in the same condition ($t_{30} = -.97$, $p = .34$). As expected, *A* is again not evaluated negatively if it does not interfere with *P*'s intrinsic tendency.

(iii) *A*-ratings in C+2 and C+3 are lower than the respective *P*-ratings ($t_{30} = -5.81$, $p < .001$, $d = -1.04$), while they do not differ significantly from the *A*-ratings in the C- conditions ($t_{30} = 1.44$, $p = .16$). Both results indicate that our subjects evaluated the concordant cases according to the same principles of opposing force-dynamics that they used in discordant cases, as predicted by our model. As soon as *A* violated the NIP, it was evaluated negatively, regardless of whether *P* was forced to reverse direction of movement or merely to continue faster on its original trajectory.

General Discussion

In this paper, we have argued that the semantic category of force dynamics (Talmy, 1988) provides a cognitive structure underlying our moral intuitions. We provided evidence that various evaluations of content-free interacting entities can be predicted from force-dynamic properties in combination with a single normative principle (NIP) that

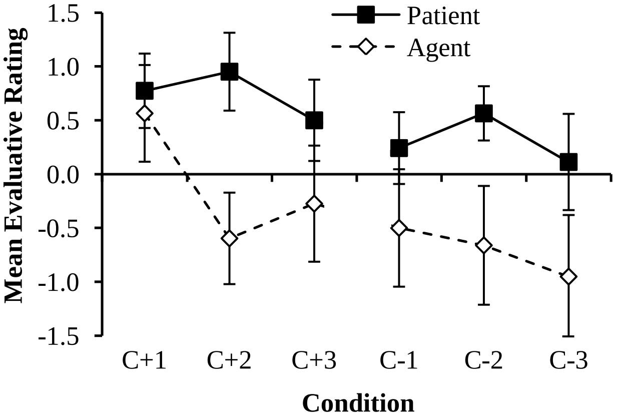

Figure 3: Results of Part 2. Error Bars = 95% CI.

expresses our prima facie moral norm not to interfere with others' interests. We thus propose that force dynamics might be part of the missing link between the apprehension of a situation and the automatic generation of a moral intuition. Observed events activate an abstract representation of the force-dynamic pattern which they instantiate. This representation is subjected to a basic normative principle, yielding default evaluative tendencies that are automatically applied to the participants of the observed interaction. We do not claim that the output of this process already represents a full-blown moral intuition. The automatically generated resulting representation could constitute a building block serving as input for higher-level processes (e.g., contextual analysis, inferences from other cues, background knowledge, application of exceptions, etc.) that eventually lead to rich, conscious moral intuitions.

Of course, the present study is only an encouraging first step in our research endeavor. So far we have only demonstrated an association between some force-dynamic patterns and explicit evaluations under maximally impoverished context conditions. It needs yet to be shown that the observed force-dynamic interactions also spontaneously elicit basic evaluations when no explicit test questions are given that request moral evaluations.

It may be seen as problematic that our model does not differentiate between animate and inanimate entities, contrary to many other theories in the field (e.g., Carey, 2009). Instead, our proposal is that force dynamic intuitions underlie event representation across domains as an abstract common code. If we are correct that basic evaluations are automatically elicited on this level of abstraction, this implies that observing one billiard ball launching another should elicit the same evaluative tendencies as observing Jack pushing Jones. Note, for example, that our displays contained no cues to animacy (such as self-propelled motion), and yet evaluations consistent with our model were observed. The postulation of a common code eliciting basic intuitions in both physical and social domains does not rule out that people use additional cues to differentiate between animate and inanimate entities (see Hamlin & Wynn, 2011, for evidence with infants). Force dynamics does not postulate that our representations of physics and psychology are *exhaustively* characterized as interplay of interacting forces. Additional semantic knowledge may of course enrich the force dynamic representation.

A related concern is that our model does not seem to capture all moral intuitions. A force-dynamic analysis of Jack lying to Jones, for example, will probably be less straightforward. We are aware that the practice of our moral judgment is very intricate and involves more considerations than those touched upon here. Our claim is thus not to provide a comprehensive theory of our moral intuitions. However, note that the range of intuitions our approach *does* potentially capture seems remarkable given its simplicity. The abstract nature of force dynamics makes it applicable to heterogeneous morally relevant events (e.g., dictators *oppressing* their people, people *resisting* temptations, etc.).

Another limitation of our approach is that it only predicts evaluations of agents. Yet, patient ratings also varied across our experimental conditions, sometimes independently from agent ratings. This might suggest that additional inferences are drawn from our stimuli which are not captured by our model.

Result (iii) of Part 2 suggests that the default conceptualization of force-related interactions is one of antagonism. It is likely that this default can quite easily be overridden if additional contextual cues are available that activate concepts of cooperation. Wolff (2007), for example, investigated cases in which *P* initially approaches a specific end state and *A* exerts a concordant force on *P*, resulting in *P* reaching the end state more quickly. Such displays reliably elicited concepts of *enable* or *help* in which *A* would presumably receive positive evaluations. We would predict that if a salient end state was provided in our displays, the default conceptualization of *P*'s intrinsic tendency might be replaced by a higher-level goal-directed conceptualization in which *P*'s intrinsic tendency would be *to reach the end state*. Once this more abstract intrinsic tendency would be attributed to *P*, the NIP would no longer be violated. Future studies will need to test these and related predictions for more complex structures with more than two protagonists.

Acknowledgements

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and by the Courant Research Centre 'Evolution of Social Behaviour', University of Göttingen (funded by the German Initiative of Excellence). We thank Laila Drummond Nauck and Julia Wiecha for helping with stimulus design and data collection.

References

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

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*, 547-619.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*, 30-39.

Michotte, A. E. (1963). *The perception of causality*. New York: Basic Books. (Original work published 1946)

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*, 49-100.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*, 580-601.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82-111.