Minimal Virtues: From Mindreading to Ethics via Joint Action (Munich)

Stephen A. Butterfill
< s.butterfill@warwick.ac.uk >

Wednesday, 8th February 2023

Contents

1. Introduction

The overall question for this talk is:

> How do agents ever perform optimally when time is pressing and cognitive resources such as working memory are scarce?

I will defend three claims:

1. New research on minmal models is needed to answer this question.
2. Pluralism about models is true.
3. Conjectures about minimal models generate readily testable predictions in the social domain.

2. A Problem for Minimal Theory of Mind

2.1. Background

Mindreading is the process of identifying a mental state as a mental state that some particular individual, another or yourself, has. To say someone has a *theory of mind* is another way of saying that she is capable of mindreading.

Butterfill & Apperly (2013) constructed a minimal theory of mind. This theory describes a model of minds and actions which, if it were implemented, would enable you to track others' false beliefs. At least within limits.

This minimal model has a signature limit: it does not enable you to track false beliefs which essentially involve a mistake about numerical identity. Such as Lois Lane's false belief that Superman and Clark Kent are distinct people (Jerry & Joe 1939).

Signature limits generate predictions. Automatic belief-tracking in adults, and belief-tracking in infants, are both subject to signature limits associated with minimal theory of mind.[1]

2.2. The Kovács Effect

Kovács et al. (2010) established that another's irrelevant belief can influence

[1] In favour: Wang et al. (2015); Low & Watts (2013); Low et al. (2014); Mozuraitis et al. (2015); Edwards & Low (2017); Fizke et al. (2017); Oktay-Gür et al. (2018); Edwards & Low (2017, 2019). Against: Kulke et al. (2018) argue that although the paradigm from Low & Watts (2013) replicates, attempts to modify it to avoid confounding factors do not produce comparable results. See also Scott et al. (2015); Carruthers (2015b,a); Kampis & Kovács (2022).

how quickly you can detect the presence of an object. Despite some initial doubts (Phillips et al. 2015), this finding has been widely replicated by several labs (including van der Wel et al. 2014; Edwards & Low 2017; El Kaddouri et al. 2020).

2.3. Question

Why do others' false beliefs ever have an effect on your own actions?

2.4. Motor Mindreading Conjecture

2.5. Predictions of the Conjecture

1. In motor mindreading only, goal-tracking will manifest sensitivity to agents' beliefs.
2. In motor mindreading only, physically constraining protagonists or participants will impair belief tracking.

This talk concerns the second prediction only.

2.6. Findings So Far

Low et al. (2020) support the prediction: physically constraining a protagonist did impair belief tracking.

Six (2022, Experiment 2) did not support the prediction: physically constraining participants did not impair their belief tracking.

And the results from a study in preparation that builds on Zani et al. (2020)'s balance paradigm found only suggestive evidence for the prediction.

3. Minimal Models for Acting Together

3.1. The Leading Theory of Acting Together (Bratman's)

Philosophical theories about joint action are based on contrasts between acting together and acting in parallel but merely individually:

> 'When we act together […] we are not each simply acting in light of expectations of the actions of others while knowing that those actions of others depend on their expectations of our actions. […] merely publicly walking alongside each other on a crowded sidewalk without colliding, while involving complex forms of mutual responsiveness, is not yet walking together in a shared

> intentional way. Can we articulate conditions that go beyond such strategic interaction and are sufficient for and illuminating of our acting together?' (Bratman 2014, pp. 1–2)

Bratman's first step towards answering this question is to postulate shared intention:

> 'A first step is to say that what distinguishes you and me from you and the Stranger is that you and I share an intention to walk together—we (you and I) intend to walk together—but you and the Stranger do not. In modest sociality, joint activity is explained by such a shared intention; whereas no such explanation is available for the combined activity of you and the Stranger. This does not, however, get us very far; for we do not yet know what a shared intention is, and how it connects up with joint action.' (Bratman 2009, p. 152)

The view that joint action involves shared intention is almost universal.

> 'I take a collective action to involve a collective [shared] intention.' (Gilbert 2006, p. 5)

> 'The sine qua non of collaborative action is a joint goal [shared intention] and a joint commitment' (Tomasello 2008, p. 181)

> 'the key property of joint action lies in its internal component [...] in the participants' having a "collective" or "shared" intention.' (Alonso 2009, pp. 444–5)

> 'Shared intentionality is the foundation upon which joint action is built.' (Carpenter 2009a, p. 381)

But what is shared intention?

Bratman's theory of shared intention has two components, a functional characterisation and a substantial 'construction of interconnected intentions and other related attitudes ... that would ... play the roles characteristic of shared intention' (Bratman 2014, p. 32).[2]

3.1.1. Bratman's Functional Characterisation

Shared intention serves to (i) coordinate activities, (ii) coordinate planning, and (iii) structure bargaining.

Bratman also proposes a requirement: shared intentions should be inferentially and normatively integrated with ordinary, individual intentions.

[2] Bratman's theory has been refined and defended over more than two decades (Bratman 1992, 1993, 1997, 2009, 2014). Here we consider just the core components.

3.1.2. Bratman's Substantial Construction

Bratman claims that the following are collectively sufficient[3] conditions for you and I to have a shared intention that we J:

> (1) '(a) I intend that we J and (b) you intend that we J
>
> (2) I intend that we J in accordance with and because of (1a), (1b), and meshing subplans of (1a) and (1b); you intend that we J in accordance with and because of (1a), (1b), and meshing subplans of (1a) and (1b)
>
> (3) (1) and (2) are common knowledge between us.' (Bratman 1993, p. View 4)

These conditions have been elaborated in later work (e.g. Bratman 2014, p. 52 on the connection condition).

3.2. Why We Need a Minimal Model

Does Bratman's account capture simple forms of joint action, such as those that appear relatively early in development? Carpenter argues that it does:

> 'I ... adopt Bratman's (1992) influential formulation of joint action or shared cooperative activity. Bratman argued that in order for an activity to be considered shared or joint each partner needs to intend to perform the joint action together ' 'in accordance with and because of meshing subplans" (p. 338) and this needs to be common knowledge between the participants' (Carpenter 2009a, p. 381).

But the hypothesis that one- and two-year-olds have shared intentions as characterised by Bratman generates a prediction: since a function of shared intention is to coordinate planning, children of this age should be capable, at least in some minimally demanding situations, of coordinating their plans with another's.

Is the prediction supported? There is good evidence that even 3-year-olds' abilities to coordinate plans are quite limited. For instance:

> '3- and 5-year-old children do not consider another person's actions in their own action planning (while showing action planning when acting alone on the apparatus). Seven-year-old children and adults however, demonstrated evidence for joint action

[3] In Bratman (1992), the following were offered as jointly sufficient *and individually necessary* conditions; the retreat to sufficient conditions occurs in Bratman (1997, pp. 143–4) where he notes that 'for all that I have said, shared intention might be multiply realizable.'

> planning. … While adult participants demonstrated the presence of joint action planning from the very first trials onward, this was not the case for the 7-year-old children who improved their performance across trials.' (Paulus 2016, p. 1059)

And:

> 'proactive planning for two individuals, even when they share a common goal, is more difficult than planning ahead solely for oneself' (Gerson et al. 2016, p. 128).

3.3. A Minimal Model of Joint Action

A goal is an outcome to which an action is directed.

An outcome is a collective goal of two or more actions involving multiple agents if it is an outcome to which those actions are collectively directed (Butterfill & Sinigaglia 2022).

Minimally, a joint action is an event involving two or more agents where the agents' actions have a collective goal.

In virtue of what do our actions have collective goals? If we answered this question by appeal to shared intention, there would be a threat of collapsing the minimal model into Bratman's model. We therefore seek an alternative answer.

One possibility is that some collective goals can be represented motorically (della Gatta et al. 2017; Sacheli et al. 2018; Clarke et al. 2019). If so, it is possible that not only intentions but also motor representations can link our actions to collective goals (Sinigaglia & Butterfill 2022).

4. Minimal Virtues in Ethical Cognition

4.1. Greene et al's Dual-Process Theory

Greene et al offer a dual-process theory of ethical cognition:

> 'this theory associates controlled cognition with utilitarian (or consequentialist) moral judgment aimed at promoting the "greater good" (Mill, 1861/1998) while associating automatic emotional responses with competing deontological judgments that are naturally justified in terms of rights or duties (Kant, 1785/1959).' (Greene 2015, p. 203)

The theory was developed in part to explain otherwise apparently anomalous responses to moral dilemmas. In particular, people have substantially

different attitudes to killing one person in order to save several others depending on whether the killing involves pressing a switch (as in the Switch dilemma) or whether it involves dropping someone through a trapdoor into the path of great danger (as in the Footbridge dilemma).[4]

What is the explanation Greene et al's theory offers?

> 'this pattern of judgment [Switch—yes; Footbridge—no] reflects the outputs of distinct and (in some cases) competing neural systems [...] The more "personal" harmful action in the footbridge case, pushing the man off the footbridge, triggers a relatively strong negative emotional response, whereas the relatively impersonal harmful action in the switch case does not.' (Greene 2015, pp. 203—4)

4.1.1. Mixed Behavioural Evidence for This Theory

One prediction of the theory is that increasing time pressure should increase the influence of automatic emotional processes relative to the influence of controlled cognition, which in turn should make responses that are characteristically deontological more likely.

This prediction is supported by (Suter & Hertwig 2011), among others.[5] But Bago & De Neys (2019) consider what happens when subjects first make a moral judgement under time pressure and extraneous cognitive load and then, just after, make another moral judgement (in answer to the same question) with no time pressure and no extraneous cognitive load. They report:

> 'Our critical finding is that although there were some instances in which deliberate correction occurred, these were the exception rather than the rule. Across the studies, results consistently showed that in the vast majority of cases in which people opt for a [consequentialist] response after deliberation, the [consequentialist] response is already given in the initial phase' (Bago & De Neys 2019, p. 1794).

Rosas & Aguilar-Pardo (2020) find, conversely to what Greene et al's theory predicts, that subjects are less likely to give characteristically deontological responses under extreme time pressure.

The converse finding of Rosas & Aguilar-Pardo (2020) is not theoretically

[4] See Greene (2015, p. 203): 'We developed this theory in response to a long-standing philosophical puzzle ... Why do people typically say "yes" to hitting the switch, but "no" to pushing?'

[5] See also Trémolière & Bonnefon (2014) and Conway & Gawronski (2013) (who manipulated cognitive load).

unmotivated—there are also some theoretical reasons for holding that automatic emotional processes should support characteristically utilitarian responses (Kurzban et al. 2012).

As there is a substantial body of neuropsychological evidence in favour of Greene et al's theory (reviewed in Greene 2014), its defenders may be little moved by the mixed behavioural evidence. But there is a reason, not decisive but substantial, to expect mixed evidence more generally …

4.1.2. Methodological Challenge

The mixed pattern of evidence for and against Greene et al's theory might be explained by their choice of vignettes using trolley cases as stimuli. Waldmann et al. (2012, p. 288) offers a brief summary of some factors which have been considered to influence responses including:

- whether an agent is part of the danger (on the trolley) or a bystander;
- whether an action involves forceful contact with a victim;
- whether an action targets an object or the victim;
- how far the agent is from the victim;[6] and
- how the victim is described.

Other factors include whether there are irrelevant alternatives (Wiegmann et al. 2020); and order of presentation (Schwitzgebel & Cushman 2015).

They comment:

> 'A brief summary of the research of the past years is that it has been shown that almost all these confounding factors influence judgments, along with a number of others [...] it seems hopeless to look for the one and only explanation of moral intuitions in dilemmas. The research suggests that various moral and nonmoral factors interact in the generation of moral judgments about dilemmas' (Waldmann et al. 2012, pp. 288, 290).

For proponents of Greene et al's view, this might be taken as encouragement. Yes, the evidence is a bit mixed. But perhaps what appears to be evidence falsifying predictions of the view will turn out to be merely a consequence of extraneous, nonmoral factors influencing judgements.

Alternatively, Waldmann et al.'s observation could be taken to suggest that few if any of the studies relying on dilemmas presented in vignette form

[6] After this review was published, Nagel & Waldmann (2013) provided substantial evidence that distance may not be a factor influencing moral intuitions after all (the impression that it does was based on confounding distance with factors typically associated with distance such as group membership and efficacy of action).

provide reliable evidence about moral factors since they do not adequately control for extraneous, nonmoral factors. As an illustration, Gawronski et al. (2017) note that aversion to killing (which would be characteristically deontological) needs to be separated from a preference for inaction. When considering only aversion to killing, time pressure appears to result in characteristically deontological responses, which would support Greene et al's theory (Conway & Gawronski 2013). But when aversion to killing and a preference for inaction are considered together, Gawronski et al. (2017) found evidence only that time pressure increases preferences for inaction.

While the combination of mixed behavioural evidence and methodological challenges associated with using dilemmas presented in vignettes does not provide a case for rejecting Greene et al's view, it does motivate considering fresh alternatives.

4.2. The Search for Minimal Virtues

I do not have a minimal model that would be useful for formulating conjectures about ethical cognition but I would like to share some ideas about where we might find one.

Step 1. Abandon the deontological/utilitarian idea with the aim, eventually, of finding a minimal model of the ethical.

Step 2. What would a minimal model be a model of? Haidt & Graham (2007) claim that there are five evolutionarily ancient, psychologically basic abilities linked to:

- harm/care
- fairness (including reciprocity)
- in-group loyalty
- respect for authority
- purity, sanctity

Step 3. Which processes might implement a minimal model? One possibility is to consider the habitual processs which support selecting goals (compare Crockett 2013 and Cushman 2013).

Habitual processes simplify the problem of goal selection by representing the world as involving only stimulus—action links. They are characterised by Thorndyke's Law of Effect:

> 'The presentation of an effective [rewarding] outcome following an action [...] reinforces a connection between the stimuli present when the action is performed and the action itself so

> that subsequent presentations of these stimuli elicit the [...] action as a response' (Dickinson 1994, p.48).

When the environment and an agent's preferences are sufficiently stable, habitual processes can approximate the computation of expected utility without the computational costs involved in identifying who probably different action outcomes are and how desirable each outcome would be (Wunderlich et al. 2012).

Habitual processes have a signature limit: they persist in extinction following devaluation.

Step 4. Find ways to interfere with habitual processes so that they are influenced by the ethical factors identified in Step 2. Initial approach: target rewards.

One possibility would be to have vicarious rewards, perhaps especially for in-group members. Suppose observing you being rewarded could trigger in me some of the reward processes that would typically occur in me if it were me, not you, who was being rewarded. Then the strength of stimulus—action links in me would be influenced not only by which outcomes are rewarding for me but also which outcomes are rewarding for you. Approximating an in-group utilitarian decision-making process.

A second possibility is inspired by aversion to bitterness as a mechanism for avoiding poisons. Poisonous foods are often bitter, and a range of animals including sea anemones become averse to a food type after a single bitter encounter (Garcia & Hankins 1975). Further, animals who encounter a greater proportion of poisonous foods in their normal diet (herbivores) show both higher sensitivity to bitterness (Li & Zhang 2014)and a higher tolerance for it (Ji 1994) .

In humans, unfairness can be detected early in the second year of life (Geraci & Surian 2011; Surian et al. 2018) and, in adults at least, unfairness can also produce sensations of bitterness.[7]

If a limited but useful range of moral violations can produce bitter sensations, general-purpose learning mechanisms can produce aversion to actions that generate these moral violations.

[7] see Chapman et al. (2009), who establish that (1) responses to bitterness are marked by activation of the levator labii muscle 'which raises the upper lip and wrinkles the nose'; (2) bitter responses are made not just to bitter tastes but also to 'photographs of uncleanliness and contamination-related disgust stimuli, including feces, injuries, insects, etc.'; and (3) in a dictator game, 'objective (facial motor) signs of disgust that were proportional to the degree of unfairness they experienced.'

5. Conclusion

Humans, even from infancy, are frequently able make good enough responses to situations when time is pressing and cognitive resources such as working memory are scarce.

Research on minimal models has provided much insight into how this is achieved in non-social domains such as physical cognition (Hubbard 2022) and animal learning (Dickinson 2016). New research on minimal models is needed to extend these successes to the social domain—to mindreading, acting together and ethical cognition.

Glossary

automatic As we use the term, a process is *automatic* just if whether or not it occurs is to a significant extent independent of your current task, motivations and intentions. To say that *mindreading is automatic* is to say that it involves only automatic processes. The term 'automatic' has been used in a variety of ways by other authors: see Moors (2014, p. 22) for a one-page overview, Moors & De Houwer (2006) for a detailed theoretical review, or Bargh (1992) for a classic and very readable introduction 2

characteristically deontological According to Greene, a judgement is *characteristically deontological* if it is one in 'favor of characteristically deontological conclusions (eg, "It's wrong despite the benefits")' (Greene 2007, p. 39). According to Gawronski et al. (2017, p. 365), 'a given judgment cannot be categorized as deontological without confirming its property of being sensitive to moral norms.' 7

collective goal an outcome to which two or more agents' actions are directed where this is not, or not only, a matter of each action being directed to that outcome (Butterfill & Sinigaglia 2022). 6

connection condition 'the condition that specifies the nature of [the] explanatory relation' between shared intention and joint action … [T]he basic idea is that what is central to the connection condition is that each is responsive to the intentions and actions of the other in ways that track the intended end of the joint action–where all this is out in the open' (Bratman 2014, pp. 78–9). 5

devaluation To *devalue* some food (or video clip, or any other thing) is to reduce its value, for example by allowing the agent to satiete themselves

on it or by causing them to associate it with an uncomfortable event such as an electric shock or mild illness. 10, 12

Drop A dilemma; also known as *Footbridge*. A runaway trolley is about to run over and kill five people. You can hit a switch that will release the bottom of a footbridge and one person will fall onto the track. The trolley will hit this person, slow down, and not hit the five people further down the track. Is it okay to hit the switch? 14

dual-process theory Any theory concerning abilities in a particular domain on which those abilities involve two or more processes which are distinct in this sense: the conditions which influence whether one mindreading process occurs differ from the conditions which influence whether another occurs. 6

extinction In some experiments, there is a phase (usually following instrumental training and an intervention such as devaluation) during which the subject encounters the training scenario exactly as it was (same stimuli, same action possibilities) but the actions produce no revant outcomes. In this extinction phase, there is no reward (nor punishment). (It is called 'extinction' because in many cases not rewarding (or punishing) the actions will eventually extinguish the stimulus–action links.) 10

Footbridge A dilemma; also known as *Drop*. A runaway trolley is about to run over and kill five people. You can hit a switch that will release the bottom of a footbridge and one person will fall onto the track. The trolley will hit this person, slow down, and not hit the five people further down the track. Is it okay to hit the switch? 7

goal A *goal* of an action is an outcome to which it is directed. 6

habitual process A process underpinning some instrumental actions which obeys *Thorndyke's Law of Effect*: 'The presentation of an effective [=rewarding] outcome following an action [...] reinforces a connection between the stimuli present when the action is performed and the action itself so that subsequent presentations of these stimuli elicit the [...] action as a response' (Dickinson 1994, p.48). (Interesting complication which you can safely ignore: there is probably much more to say about under what conditions the stimulus–action connection is strengthened; e.g. Thrailkill et al. 2018.) 9

joint action Many of the things we do are, or could be, done with others. Mundane examples favoured by philosophers include painting a house

together (Bratman 1992), lifting a heavy sofa together (Velleman 1997), preparing a hollandaise sauce together (Searle 1990), going to Chicago together (Kutz 2000), and walking together (Gilbert 1990). These examples are supposed to be paradigm cases of a class of phenomena we shall call 'joint actions'.

Researchers have used a variety of labels including 'joint action' (Brooks 1981; Sebanz et al. 2006; Knoblich et al. 2011; Tollefsen 2005; Pettit & Schweikard 2006; Carpenter 2009b; Pacherie 2010; Brownell 2011; Sacheli et al. 2018; Meyer et al. 2013), 'social action' (Tuomela & Miller 1985), 'collective action' (Searle 1990; Gilbert 2010), 'joint activity' (Baier 1997), 'acting together' (Tuomela 2000), 'shared intentional activity' (Bratman 1997), 'plural action' (Schmid 2008), 'joint agency' (Pacherie 2013), 'small scale shared agency' (Bratman 2014), 'intentional joint action' (Blomberg 2016), 'collective intentional behavior' (Ludwig 2016), and 'collective activity' (Longworth 2019).

We leave open whether these are all labels for a single phenomenon or whether different researchers are targeting different things. As we use 'joint action', the term applies to everything any of these labels applies to. 3

meshing subplans 'The sub-plans of the participants *mesh* when it is possible that all of these sub-plans taken to gether be successfully executed.' (Bratman 2014, p. 53) 5

model A model is a way some part or aspect of the world could be. 2

modest sociality 'small scale shared intentional agency in the absence of asymmetric authority relations' (Bratman 2009, p. 150). 4

outcome An outcome of an action is a possible or actual state of affairs. 11

shared intention An attitude that stands to joint action as ordinary, individual intention stands to ordinary, individual action. It is hard to find consensus on what shared intention is, but most agree that it is neither shared nor intention. (Variously called 'collective', 'we-' and 'joint' intention.) 4, 6

signature limit A *signature limit* of a system is a pattern of behaviour the system exhibits which is both defective given what the system is for and peculiar to that system. A *signature limit* of a model is a set of predictions derivable from the model which are incorrect, and which are not predictions of other models under consideration. 2, 10

Switch A dilemma; also known as *Trolley*. A runaway trolley is about to run over and kill five people. You can hit a switch that will divert the trolley onto a different set of tracks where it will kill only one. Is it okay to hit the switch? 7

track For a process to *track* an attribute or thing is for the presence or absence of the attribute or thing to make a difference to how the process unfolds, where this is not an accident. (And for a system or device to track an attribute is for some process in that system or device to track it.)

Tracking an attribute or thing is contrasted with *computing* it. Unlike tracking, computing typically requires that the attribute be represented. 2

Transplant A dilemma. Five people are going to die but you can save them all by cutting up one healthy person and distributing her organs. Is it ok to cut her up? 14

Trolley A dilemma; also known as *Switch*. A runaway trolley is about to run over and kill five people. You can hit a switch that will divert the trolley onto a different set of tracks where it will kill only one. Is it okay to hit the switch? 14

trolley cases Scenarios designed to elicit puzzling or informative patterns of judgement about how someone should act. Examples include Trolley, Transplant, and Drop. Their use was pioneered by Foot (1967) and Thomson (1976), who aimed to use them to understand ethical considerations around abortion and euthanasia. 8

References

Alonso, F. M. (2009). Shared intention, reliance, and interpersonal obligations. *Ethics*, *119*(3), 444–475.

Bago, B. & De Neys, W. (2019). The Intuitive Greater Good: Testing the Corrective Dual Process Model of Moral Cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801.

Baier, A. C. (1997). Doing Things With Others: The Mental Commons. In L. Alanen, S. Heinamaa, & T. Wallgren (Eds.), *Commonality and particularity in ethics* (pp. 15–44). Palgrave Macmillan.

Bargh, J. A. (1992). The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects. *The American Journal of Psychology*, *105*(2), 181–199.

Blomberg, O. (2016). Common Knowledge and Reductionism about Shared Agency. *Australasian Journal of Philosophy*, *94*(2), 315–326.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Bratman, M. E. (1993). Shared intention. *Ethics*, *104*, 97–113.

Bratman, M. E. (1997). I intend that we J. In R. Tuomela & G. Holmstrom-Hintikka (Eds.), *Contemporary Action Theory, Volume 2: Social Action*. Dordrecht: Kluwer. Reprinted in Bratman, M. (1999) *Faces of Intention*. Cambridge: Cambridge University Press (pp. 142-161).

Bratman, M. E. (2009). Modest sociality and the distinctiveness of intention. *Philosophical Studies*, *144*(1), 149–165.

Bratman, M. E. (2014). *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Brooks, D. H. M. (1981). Joint action. *Mind*, *90*(357), 113–119.

Brownell, C. A. (2011). Early Developments in Joint Action. *Review of Philosophy and Psychology*, *2*, 193–211.

Butterfill, S. A. & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*(5), 606–637.

Butterfill, S. A. & Sinigaglia, C. (2022). Towards a Mechanistically Neutral Account of Acting Jointly: The Notion of a Collective Goal. *Mind*, *X*(X), fzab096.

Carpenter, M. (2009a). Just how joint is joint action in infancy? *Topics in Cognitive Science*, *1*(2), 380–392.

Carpenter, M. (2009b). Just how joint is joint action in infancy? *Topics in Cognitive Science*, *1*(2), 380–392.

Carruthers, P. (2015a). Mindreading in adults: evaluating two-systems views. *Synthese*, *forthcoming*, 1–16.

Carruthers, P. (2015b). Two systems for mindreading? *Review of Philosophy and Psychology*, *7*(1), 141–162.

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In Bad Taste: Evidence for the Oral Origins of Moral Disgust. *Science*, *323*(5918), 1222–1226.

Clarke, S., McEllin, L., Francová, A., Székely, M., Butterfill, S. A., & Michael, J. (2019). Joint action goals reduce visuomotor interference effects from a partner's incongruent actions. *Scientific Reports*, *9*(1), 1–9.

Conway, P. & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of personality and social psychology*, *104*(2), 216–235.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

della Gatta, F., Garbarini, F., Rabuffetti, M., Viganò, L., Butterfill, S. A., & Sinigaglia, C. (2017). Drawn together: When motor representations ground joint actions. *Cognition*, *165*, 53–60.

Dickinson, A. (1994). Instrumental conditioning. In N. Mackintosh (Ed.), *Animal Learning and Cognition*. London: Academic Press.

Dickinson, A. (2016). Instrumental conditioning revisited: Updating dual-process theory. In J. B. Trobalon & V. D. Chamizo (Eds.), *Associative learning and cognition*, volume 51 (pp. 177–195). Edicions Universitat Barcelona.

Edwards, K. & Low, J. (2017). Reaction time profiles of adults' action prediction reveal two mindreading systems. *Cognition*, *160*, 1–16.

Edwards, K. & Low, J. (2019). Level 2 perspective-taking distinguishes automatic and non-automatic belief-tracking. *Cognition*, *193*, 104017.

El Kaddouri, R., Bardi, L., De Bremaeker, D., Brass, M., & Wiersema, J. R. (2020). Measuring spontaneous mentalizing with a ball detection task: Putting the attention-check hypothesis by Phillips and colleagues (2015) to the test. *Psychological Research*, *84*(x), 1749–1757.

Fizke, E., Butterfill, S. A., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Signature limits in early theory of mind: Toddlers spontaneously take into account false beliefs about an objects' location but not about its identity. *Journal of Experimental Child Psychology*, *forthcoming*.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Garcia, J. & Hankins, W. (1975). *The Evolution of Bitter and the Acquisition of Toxiphobia*, (pp. 39–45). Elsevier.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of personality and social psychology*, *113*(3), 343–376.

Geraci, A. & Surian, L. (2011). The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources. *Developmental Science*, *14*(5), 1012–1020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2011.01048.x.

Gerson, S. A., Bekkering, H., & Hunnius, S. (2016). Social context influences planning ahead in three-year-olds. *Cognitive Development*, *40*, 120–131.

Gilbert, M. P. (1990). Walking together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy*, *15*, 1–14.

Gilbert, M. P. (2006). Rationality in collective action. *Philosophy of the Social Sciences*, *36*(1), 3–17.

Gilbert, M. P. (2010). Collective action. In T. O'Connor & C. Sandis (Eds.), *A Companion to the Philosophy of Action* (pp. 67–73). Oxford: Blackwell.

Greene, J. D. (2007). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3* (pp. 35–79). MIT Press.

Greene, J. D. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726.

Greene, J. D. (2015). The cognitive neuroscience of moral judgment and decision making. In *The Moral Brain: A Multidisciplinary Perspective* (pp. 197–220). Cambridge, MA, US: MIT Press.

Haidt, J. & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116.

Hubbard, T. L. (2022). The possibility of an impetus heuristic. *Psychonomic Bulletin & Review*, *29*(6), 2015–2033.

Jerry, S. & Joe, S. (1939). Superman. *Superman.*, (1). OCLC: 5911080.

Ji, G. (1994). Is the bitter rejection response always adaptive? *Physiology & Behavior*, *56*(6).

Kampis, D. & Kovács, Á. M. (2022). Seeing the World From Others' Perspective: 14-Month-Olds Show Altercentric Modulation Effects by Others' Beliefs. *Open Mind*, *5*, 189–207.

Knoblich, G., Butterfill, S. A., & Sebanz, N. (2011). Psychological research on joint action: Theory and data. In B. Ross (Ed.), *Psychology of Learning and Motivation*, volume 51 (pp. 59–101). San Diego, CA: Academic Press.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830 –1834.

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study. *Psychological Science*, 0956797617747090.

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior*, *33*(4), 323–333.

Kutz, C. (2000). Acting together. *Philosophy and Phenomenological Research*, *61*(1), 1–31.

Li, D. & Zhang, J. (2014). Diet Shapes the Evolution of the Vertebrate Bitter Taste Receptor Gene Repertoire. *Molecular Biology and Evolution*, *31*(2), 303–309.

Longworth, G. (2019). Sharing non-observational knowledge. *Inquiry*, *0*(0), 1–21.

Low, J., Drummond, W., Walmsley, A., & Wang, B. (2014). Representing how rabbits quack and competitors act: Limits on preschoolers' efficient ability to track perspective. *Child Development*, *forthcoming*.

Low, J., Edwards, K., & Butterfill, S. A. (2020). Visibly constraining an agent modulates observers' automatic false-belief tracking. *Scientific Reports*, *10*(1), 11311.

Low, J. & Watts, J. (2013). Attributing false-beliefs about object identity is a signature blindspot in humans' efficient mindreading system. *Psychological Science*, *24*(3), 305–311.

Ludwig, K. (2016). *From Individual to Plural Agency: Collective Action*. Oxford University Press.

Meyer, M., van der Wel, R. P. R. D., & Hunnius, S. (2013). Higher-order action planning for individual and joint object manipulations. *Experimental Brain Research*, *225*(4), 579–588.

Moors, A. (2014). Examining the mapping problem in dual process models. In *Dual process theories of the social mind* (pp. 20–34). Guilford.

Moors, A. & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326.

Mozuraitis, M., Chambers, C. G., & Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition*, *142*, 148–165.

Nagel, J. & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 237.

Oktay-Gür, N., Schulz, A., & Rakoczy, H. (2018). Children exhibit different performance patterns in explicit and implicit theory of mind tasks. *Cognition*, *173*, 60–74.

Pacherie, E. (2010). The phenomenology of joint action: Self-agency vs. joint-agency. In A. Seemann (Ed.), *Joint Action*. MIT Press.

Pacherie, E. (2013). Intentional joint agency: shared intention lite. *Synthese*, *190*(10), 1817–1839.

Paulus, M. (2016). The development of action planning in a joint action context. *Developmental Psychology*, *52*(7), 1052–1063.

Pettit, P. & Schweikard, D. (2006). Joint Actions and Group Agents. *Philosophy of the Social Sciences*, *36*(1), 18 –39.

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*(9), 1353–1367.

Rosas, A. & Aguilar-Pardo, D. (2020). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning*, *26*(4), 534–551.

Sacheli, L. M., Arcangeli, E., & Paulesu, E. (2018). Evidence for a dyadic motor plan in joint action. *Scientific Reports*, *8*(1), 5027.

Schmid, H. B. (2008). Plural action. *Philosophy of the Social Sciences*, *38*(1), 25–54.

Schwitzgebel, E. & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127–137.

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32–56.

Searle, J. R. (1990). Collective intentions and actions. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in Communication* (pp. 90–105). Cambridge: Cambridge University Press. Reprinted in Searle, J. R. (2002) *Consciousness and Language*. Cambridge: Cambridge University Press (pp. 90–105).

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and mind moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76.

Sinigaglia, C. & Butterfill, S. A. (2022). Motor representation in acting together. *Synthese*, *200*(2), 82.

Six, P. (2022). *Observing Action in Uncertainty: The Role of Belief-tracking in Action Observation*. Thesis, Te Herenga Waka-Victoria University of Wellington.

Surian, L., Ueno, M., Itakura, S., & Meristo, M. (2018). Do Infants Attribute Moral Traits? Fourteen-Month-Olds' Expectations of Fairness Are Affected by Agents' Antisocial Actions. *Frontiers in Psychology*, *9*.

Suter, R. S. & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458.

Thomson, J. J. (1976). Killing, Letting Die, and The Trolley Problem. *The Monist*, *59*(2), 204–217.

Thrailkill, E. A., Trask, S., Vidal, P., Alcalá, J. A., & Bouton, M. E. (2018). Stimulus control of actions and habits: A role for reinforcer predictability and attention in the development of habitual behavior. *Journal of Experimental Psychology: Animal Learning and Cognition*, *44*, 370–384.

Tollefsen, D. (2005). Let's pretend: Children and joint action. *Philosophy of the Social Sciences*, *35*(75), 74–97.

Tomasello, M. (2008). *Origins of human communication*. The MIT Press.

Trémolière, B. & Bonnefon, J.-F. (2014). Efficient Kill–Save Ratios Ease Up the Cognitive Demands on Counterintuitive Moral Utilitarianism. *Personality and Social Psychology Bulletin*, *124*(3), 379–384.

Tuomela, R. (2000). *Cooperation: A Philosophical Study*. Dordrecht: Springer.

Tuomela, R. & Miller, K. (1985). We-Intentions and Social Action. *Analyse & Kritik*, *7*(1), 26–43.

van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? evidence from a continuous measure. *Cognition*, *130*(1), 128–133.

Velleman, D. (1997). How to share an intention. *Philosophy and Phenomenological Research*, *57*(1), 29–50.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral Judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 274–299). Oxford: Oxford University Press.

Wang, B., Hadi, N. S. A., & Low, J. (2015). Limits on efficient human mindreading: Convergence across chinese adults and semai children. *British Journal of Psychology*, *106*(4), 724–740.

Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive expertise and irrelevant options. *Oxford studies in experimental philosophy*, *3*, 275–310.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786–791.

Zani, G., Butterfill, S. A., & Low, J. (2020). Mindreading in the balance: Adults' mediolateral leaning and anticipatory looking foretell others' action preparation in a false-belief interactive task. *Royal Society Open Science*, *7*(1), 191167.