Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Joint International PhD Program in Cognitive Neuroscience

Ciclo XXX

Settore Concorsuale: 11/E1

Settore Scientifico Disciplinare: M-PSI/01

Does intentionality decision-making depend on who you are? The role of individual differences

Presentata da: **Dott.ssa Micaela Maria Zucchelli**

Coordinatore Dottorato	Supervisore
Prof.ssa Monica Rubini	**Prof.ssa Fiorella Giusberti**

Esame finale anno 2018

To Rita and Giandomenico, who first inspired me to travel this road,

And to Marco who, like them, is always by my side.

"*Don't let us forget that the causes of human actions are usually immeasurably more complex and varied than our subsequent explanations of them*." (Fëdor Dostoevskij)

Table of Contents

Abstract

Intentionality attribution is a critical ability in everyday life, necessary for attributing meaning to others' actions. Any impairment in its ascription has been shown to produce significant difficulties in the handling of one's social life. Recent evidence has indicated that healthy individuals may exhibit systematic bias in their assessment of other people's intentions: when presented with the negative side effect of an action, the majority of people tend to judge it as intentionally performed, as opposed to a positive side effect (*Knobe Effect*).Recent research has considered the role of the individual characteristics of the person who judges in order to explain this effect.

The experiments included in this dissertation aim to explore the role of certain individual differences strongly associated with intentionality attribution. Specifically, Study 1 revealed that individual production of downward counterfactuals decreases intentionality attributions, whereas upward ones increase them. Study 2 explored the role of individual differences in Theory of Mind (ToM) ability, finding that higher ToM allows individuals to focus on information about the intentions of the agent, reducing attention towards side effects and thus the *Knobe Effect*. Study 3 confirmed this result, showing that individuals with autistic personality traits tend to over-attribute intentionality to the side effects, due to their reduced attention to intentions. Finally, Study 4 explored the influence of individual ability in processing emotions, by testing individuals with alexithymic personality traits who, conversely, are less influenced by the side effect information and reduce the intentionality attributed to them.

In summary, according to their individual characteristics, people focus their attention on different elements (e.g intentions-side effects) while analyzing social situations, which are consequently perceived in different ways. Future conceptual models of intentionality should take into greater account the influence of individual differences in determining which elements people focus on when ascribing meaning to others' actions.

Chapter 1. General introduction

1.1 Intentionality: a definition

"Intentionality is a core category of mental life, along with space, time and cause" (Miller & Johnson-Laird, 1976). The attribution of intentionality represents a crucial decision-making process, largely analyzed by several disciplines such as philosophy, psychology and law (e.g. Miller & Jonson–Laird, 1976; Malle & Knobe, 1997; Malle & Nelson, 2003).

Possessing an intention has strong implications because it involves a responsibility for the action by the agent who performs it. Indeed, when perceiving, explaining, or criticizing human behaviour, people generally distinguish between intentional and unintentional actions: intentional actions are typically interpreted as deliberate and explained on the basis of reasons, beliefs and desires; whereas unintentional behaviours are considered to be random and explained more by causes which are outside of the person.

This distinction is especially decisive in the legal system where intentionality plays a crucial role, establishing the difference between intentional murder and manslaughter, and through its close ties with assessments of responsibility and blame (e.g. Malle & Nelson, 2003). Indeed, the attribution of responsibility to the agent closely depends on his/her mental states when carrying out the action, and in particular on whether he intentionally produced a given outcome (Buon, Jacob, Loissel, & Dupoux, 2013; Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007).

At large, intentionality attribution represents a key capability in the everyday life of individuals, in which we constantly need to ascribe meaning to other people's actions, in order to respond to them with the appropriate behaviour. Therefore it permeates and sets the course of all our social interactions.

Moreover, according to Rossett (2008) there is a tendency, between individuals, to over-attribute intention when presented with ambiguous actions, that could be interpreted as either intentional or unintentional. This mechanism is supposed to represent an adaptive cognitive heuristic which allows us to act quickly, for instance if a negative event occurs (Moore & Pope, 2014), as, according to the author, the risk of a false positive error (reasoning that an action was intentional when it was instead an accident) is lower than that of a false-negative error (reasoning that an action was accidental when it was instead intentional). However, the present dissertation will raise awareness of the former risk.

The importance of capability to attribute intentionality is also confirmed by its early development within human abilities: during their first year of life, infants showed they could distinguish goal-directed human actions from other events (e.g., Premack, 1990; Sommerville & Woodward, 2005; Wellman & Phillips, 2001; Gergely, Nádasdy, Csibra, & Bíró, 1995) and three year-olds are able to assign more blame for intentional than accidental behaviour (e.g., Nunez & Harris,1998).

Therefore, it is fundamental to establish when a behaviour can be considered intentional. Several researchers have tried to answer this question, however, the multidisciplinary nature of this concept has made it difficult to study and categorize empirically.

Traditionally, psychology and philosophy have studied intentionality as an “objective fact about the mind” (e.g. Fishbein & Ajzen, 1975; Heckhausen, 1991; Libet, 1985; Ryan, 1970; Schneider & Shiffrin, 1977); however since people ascribe intentions to each other, the social role of this concept has led researchers to investigate whether a shared folk concept of intentionality could exist, and which specific features characterize an action as intentional.

Malle and Knobe (1997) were the first who empirically explored the folk concept of intentionality, whereas previously this concept had been studied in relation to its influence on other psychological processes, such as attribution of responsibility, guilt or punishment, but not with regards to the concept of intentionality itself (e.g., Piaget, 1932; Shaver, 1985; Shultz

& Wright, 1985). Moreover, some previous attempts were made about the characteristics of this concept but without reaching a shared definition: these theoretical models disagreed on what they identified as the necessary conditions for intentional actions:

Heider's (1958) model of intentional action recognized the "intention" and the "desire" to do the action as components of intentionality concept, but overlooked the role of "beliefs" about the action. Jones and Davis (1965) extended Heider's model by recognizing the importance of a "belief" component (also called ''knowledge''), in addition to the component "ability" to carry out the action, as necessary conditions for ascriptions of intention whereas the role of the "desire" component remained unspecified. Ossorio and Davis (1968) postulated the importance of "desire", "knowledge" (meaning belief), and "skill"(meaning the ability to carry out the action) components (p. 358). Later, Shaver (1985) enriched the way to define intentional action by claiming that an action is intentional if the agent has a "desire for an outcome", "beliefs about the consequences of the action", and "beliefs about his or her ability to perform the action". Importantly, this definition includes the "desire" component but merges "intention" with "intentional action" and supplants "ability" with "perceptions of ability". Finally, Forguson (1989) claimed that to act intentionally, one only needs to have a "desire (for an outcome)" and appropriate "beliefs (about how the act would lead to that outcome)".

Despite a large investigation into the intentionality concept, neither of these models have been based on empirical data, whereas Malle and Knobe (1997) explored the concept of "intentionality" among people. In their study, the authors asked the participants to provide definitions for the concept of intentionality, using the following question: ''*When you say that somebody performed an action intentionally, what does this mean? Please explain*''. By coding the answers provided, the authors highlighted that individuals possess a shared folk concept of intentionality which includes, in addition to the classic components such as "desire", "belief" and "intention", a new component: the "awareness of performing an act" (the agent's state of mind at the time of acting). Moreover the definitions provided also outlined an important difference between the "intentions" and "desires" concepts: intentions always have an action as their object , whereas desires can take any outcome as their object (even impossible states of the world). Therefore the model comprised: (a) a *desire* for an outcome, (b) *beliefs* about the action leading to that outcome, (c) an *intention* to perform that act, and (d) *awareness* of performing that act; whereas the *skill (or ability)* component, postulated in previous models, was absent from people's definitions. According to the authors, participants failed to mention *skill* in their definitions of intentionality because they only considered interpersonal behaviours for which skill can be assumed. Therefore, in a subsequent study, they provided participants with two scenarios focused on the "skill" component. In the example, the authors compared a situation in which an agent who had never played darts before, surprisingly achieves a high score in his first try (by hitting the central part of the target in which the points are tripled) whereas he failed trying again, to a situation in which the same individual obtained the same high score twice in a row, so apparently proving more skill than in the first one which seemed to occur by chance. Indeed, participants only judged the latter condition to be intentional, showing that they considered an action intentional if there is evidence of the agent's skill. So, the agent in the first condition had the *intention* to hit the central part of the target (he tried to hit it), but he did not hit it *intentionally*, because judgments of intentionality require evidence of skill in addition to evidence of intention.[1] In this way the authors clarified a terminology complexity between the concept of *intention* and *intentional* actions, often considered synonyms (e.g., Jones & Davis, 1965; Shaver, 1985). According to their evidence, people distinguish between attributions of *intention* (based on belief and desire) and attributions of doing something *intentionally* (based

[1] However, in a later article, Knobe (2003b) rejected the view that skill is always essential, finding that evaluative considerations have a significant influence on whether people consider skill to be crucial for acting intentionally; for instance, people tend to judge immoral actions to be intentional, even when the agent has demonstrably no skill.

on intention, skill, and awareness), and apply the term "intention" to persons (who intend to do something) and the term "intentional" to actions (which are performed intentionally).

This distinction also revealed a hierarchical relationship between the components of the folk concept of intentionality: if an action lacks awareness and skill it can be considered *intended* but not performed *intentionally*, whereas a person cannot intend a specific action, if s/he does not have the desire and belief behind that action. (Figure 1)

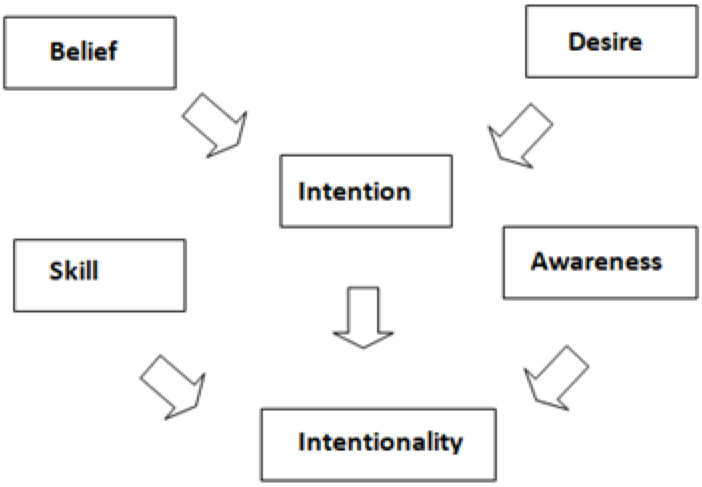

Figure 1. The folk concept of intentionality model from Malle and Knobe (1997).

More recently, Skulmowski et al. (2015) criticized the methodology employed by Malle and Knobe (1997), claiming that their methodology is incapable of exhaustively explaining the notion of intentionality. The authors agreed with Malle and Knobe's general approach of employing open questions but they considered the classifications adopted too strongly driven by the researchers' preconceptions, rather than by the participants' actual responses. Conversely, they asked participants to construct their own scenarios and give their own explanations for why they thought that a given action was intentional or unintentional.

According to them, this procedure may help people in providing their definitions because "*the notion of intentional action is too abstract and giving concrete examples may facilitate the recall of important information that would otherwise be rarely available*".

The study detected "intentions, desires, decisions, and thoughts about actions" as components belonging to the intentional concept, whereas the "belief" component was not explicitly cited, but conceptually implied by other categories. Moreover, people's conception of unintentional action was not only an inversion of their conception of intentional action since, in addition "to lack of intention and lack of desire", this concept is strongly linked to "inattention, lack of control, and accidents". Finally, the study revealed that people predominantly associate unintentional actions with bad outcomes, and link intentional actions more strongly to positive outcomes. The authors concluded claiming that different approaches would work better in conjunction, in order to provide a clear definition of the concept of intentionality.

Later, Moss (2017) underlined that this approach suffers from two major limitations, as well: i) it lacks the opportunity for researchers to follow-up and dialogue with participants to confirm their understanding; ii) the analysis focuses on taking qualitative content but coding it quantitatively, thus losing some individual details.

1.2 Bias and loss of capability of intentionality attribution

Regardless of the methodology used, determining the intentionality of an action is not always straightforward. A prototypical example is highlighted by Knobe (2003a), who showed that, in some circumstances, a behaviour can be considered intentional even if it is unintended but foreseen as a side effect.

The employed scenario describes a situation in which the vice-president of a company proposes a new program to the Chairman, which would increase the company's profits. This program involves two possible side-effects: in the first scenario, the profit growth will also bring about harmful effects to the environment, while in the second case the profits will also produce beneficial effects. In both conditions the Chairman asserts that he doesn’t care about the side effects since he just wants to make as much profit as he can.

Here are the two versions of the classic scenario:

> **Harm**
>
> *The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.” The chairman of the board answered, “I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.” They started the new program. Sure enough, the environment was harmed.*
>
> **Help**
>
> *The vice-president of the company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.” The chairman of the board answered, “I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.” They started the new program. Sure enough, the environment was helped.*

(Knobe, 2003a, p. 191).

Participants had to determine whether the Chairman intentionally harmed or helped the environment. Surprisingly, the two conditions elicited two radically different patterns of responses: in the harm condition most participants (82%) considered that the Chairman caused the environmental damage intentionally, while in the help condition, only 23% considered the environmental benefits performed intentionally. These results have been replicated with a variety of scenarios (Knobe, 2003a, 2003b, 2004a, 2004b; Knobe & Mendlow, 2004; Mele & Cushman, 2007; Nadelhoffer, 2004a, 2004b, 2006), and the effect has proven to be robust across different cultures (Knobe & Burra, 2006, but see Robbins et al., 2017), and ages (Leslie, Knobe, & Cohen, 2006; Pellizzoni et al.,2009). This asymmetry in intentionality attribution has become known as “the *Knobe Effect*” (*KE*).

According to the common view of classic theories, originating from Tommaso D'Aquino's *Doctrine of Double Effect* to contemporary theories (e.g. Mikhail, 2007), attributions of intentionality represent an input for moral evaluations: an action's value must be judged based on the intentions that caused it and on the lack of alternatives that would prevent the occurrence of side effects (that do not have to be incommensurate to the results achieved with the action) whereas the value of the effects of an action are irrelevant to the judgments of intentionality. Therefore whether or not an action is done intentionally influences judgments about moral wrongness or rightness and not the reverse (e.g., Heider 1958; Schlenker et al. 1994).

In contrast the described *KE* overturned classic theories suggesting that certain kinds of moral evaluations might actually shape whether people interpret an agent's behavior to be intentional or not (Cova et al., 2016), and it has opened up an intense debate on its nature and possible implications for the folk notion of intentionality (e.g., Knobe 2006; Machery 2008; Nichols & Ulatowski, 2007).

Moreover, these findings led researchers to particular concerns about the risks of biased decisions and reduced impartiality from jurors called to attribute intentionality, who could be influenced by the moral value of the action considered (Adams, 2015; Nadelhoffer, 2006); whereas others concluded that Knobe's results outlined the existence of a gap between the law and the way we ordinarily think about responsibility, and argued the law should be changed accordingly (Duff, 2015; Kobick, 2010).

In relation to the concern about the risks of a reduced impartiality from jurors, Malle and Nelson (2003, p. 578) suggested three steps to undertake in order to reduce these concerns: "*The first is to recognize and document the systematic and consistent nature of folk concepts of the mind that people apply in ordinary life. The second is to select those folk concepts that are important for the law (such as the components of intentionality) and use them in their original meaning when formulating laws as well as jury instructions. Technical concepts may* *be introduced as well, but their labels must be sufficiently distinct from folk concepts to avert interference and confusion. Finally, if the selected concepts of mind are to serve a distinct function in the law, they must be separated from notions of culpability and punishment. Similarly, jurors' judgments of intention and other mental states must be dissociated from evaluative feelings of anger, outrage, and vengeance.*"

Understanding that common people make asymmetries in judging the intentionality of others' actions is crucial, especially when we consider the practical consequences of this important decision-making process in everyday social interactions.

As stated previously, understanding others' intentions is fundamental to establish and maintain social relationships. A diminished ability to understand intentions, such as not understanding the innocent mistakes of other people, can lead to the interruption of social interactions or relationships.

On the contrary, the absence of this capacity could lead to catastrophic consequences such as:

(i) in patients with a ventromedial prefrontal lesion, who tend to judge attempted harm to others, including attempted murder, as more morally permissible, compared to the control group. Indeed these individuals, due to their lesion, tend to neglect the protagonist's negative intention, focusing instead on the action's neutral outcome (e.g., Young et al., 2010; Ciaramelli et al., 2012);

(ii) in patients with a behavioral variant of frontotemporal dementia who, similarly to patients affected by ventromedial lesions, consider attempted harm more morally permissible, and furthermore they are less willing to exonerate individuals for accidentally causing harm. Indeed, exculpating an agent who causes harm accidentally requires an especially robust representation of the intentions (neutral, in this case), as it is necessary to use this information to override the preponderant negative effect produced by the negative outcome (e.g. Baez et al., 2014);

(iii) in schizophrenic patients, who exhibit delusions of reference or beliefs of conspiracy against them, due to the misunderstanding of others' intentions (Sarfati et al.,1997; Brüne, 2005; Fletcher & Frith, 2009).

1.3 Interpretations of the Knobe Effect

The experiment performed by Knobe (2003a) revealed that our intentionality attribution can be influenced by certain factors. The following are some of the major interpretations of *KE*. First of all a main distinction has been postulated between Knobe’s interpretation and that of other researchers (Hindriks et al., 2016): Knobe believes that the asymmetry in the attributions of intentionality pertains to the folk concept of intentionality existing among people, whereas many other authors hold that the *KE* is due to external factors that interfere with those conceptual competences, for instance a responsibility on the part of the agent for the side effects produced.

1.3.1 The moral valence hypothesis

The first interpretation provided by Knobe (2006) suggested that a prior morality judgment, about the negative outcome, could influence the subsequent intentionality attribution: in particular, when people judge the intentionality of an action that has moral consequences, they fail to consider some of the components of intentionality, such as the agent’s skill and the desire for the action, and “are quite likely to consider the immoral action intentional even if, by strict standards, it may not be intentional” (Malle, 2006). According to Knobe (2006), the folk concept of intentional action is sensitive to the moral valence of the side effects: first, one determines if the behaviour is good or bad and then one searches for sufficient features of the negative side effect to be able to judge it intentionally performed.

Against this proposal however, it has been observed that people judge a protagonist to have brought about a harmful side-effect intentionally even for non-moral side-effects. For instance Phelan and Sarkissian (2008) provided participants with scenarios describing a president of a corporation who intends to increase sales in Massachusetts and foresees, but doesn't care, that it will decrease sales in New Jersey. Most people judged that the president lowered sales in New Jersey intentionally, but unlike the classic *KE*, most judged the side effect not to be bad.

On the other hand, some side effects can be judged unintentional even though they are judged to be bad, for instance a town-planner who introduces a programme to clean toxic waste, with the side effect of increasing unemployment, is judged to have affected unemployment unintentionally even though enhancing unemployment is judged to be bad (e.g., Phelan & Sarkissian, 2008).

Therefore the moral valence of the side effect does not seem to fully explain the *KE*.

1.3.2 The Conversational Hypothesis

Adams and Steadman (2004a, 2004b) proposed that the *KE* is produced by conversational implicature. They argue that in the harm condition, participants want to express their blame towards the Chairman, therefore they affirmed that he intentionally performed the side effect. Otherwise if participants say that the chairman did not bring about the negative side effect intentionally, then that would conversationally imply that the chairman is not blameworthy of it.

Against this view, however, if participants are allowed to express their ratings of blame, so there is no need to conversationally imply that the chairman intentionally performed the harmful side effect, the asymmetry persists (Malle, 2006).

Similarly, Guglielmo and Malle (2010) explored linguistic implicatures involved in the Effect, giving participants the chance to choose the most appropriate description of the Chairman's action, from among the following four:

"*The chairman willingly harmed the environment*", "*The chairman knowingly harmed the environment*", "*The chairman intentionally harmed the environment*" and "*The chairman purposefully harmed the environment".* They found that few participants still chose the description which stated that the Chairman intentionally harmed the environment. However, this result does not imply that, if participants consider a claim not to be the most accurate, they also consider it to be false (Cova et al., 2016).

1.3.3 The Blame and Affective Bias Hypothesis

One of the most influential explanations of the *KE* explored, instead, the role of emotions and blame in intentionality attribution, through different approaches, starting from Alicke's Culpable Control Model of blame (CCM) (Alicke, 2000). This model posited that the tendency of people to attribute blame makes them construe evidence in a way that supports the attribution they want to make. In the specific Harm case of the chairman vignette, negative evaluations of the Chairman's attitude and the outcome of his action lead people to attribute intentionality in order to justify blame (Alicke & Rose, 2012). According to the authors, people blame first and search for mitigating circumstances later.

In line with Alicke's model, Nadelhoffer (2004b) argued that the perceived blameworthiness of the Chairman influences the asymmetry, because the harm and help situations are not analogous. In fact, in both scenarios, the Chairman does not care about something he should care about, producing opposing effects: people do not want to praise him for the help side-effects, whereas they want to blame him for the harmful effect. So the harmful condition is seen as more blameworthy, compared to how praiseworthy the helpful one is seen to be.

Nadelhoffer stated that if the help condition was perceived as praiseworthy, people would judge the positive side effect intentionally as well. Therefore he proposed a new scenario in which two friends are competing against each other in an essay contest: *Jason helps edit his friend's essay, unconcerned that in doing so, he will lower his own chances of winning the* *contest; and Jason in fact does not win.* In this case most people (55%) judged that Jason does decrease his chances intentionally and is praiseworthy for doing so, eliminating the *KE* (Nadelhoffer, 2004b, p. 210).

Moreover, affective processes can increase the motivation to blame, and thus represent an extension of the motivational bias (Diaz et al., 2017). The role of emotions in increasing intentionality judgments to the negative side effect is supported by Malle and Nelson (2003), who argued that since one has a negative affective reaction to the Chairman in the harm scenario, our judgments about intentionality are biased and so one is more likely to think that the harm is brought about intentionally.

However, this hypothesis was refused by Young et al. (2006), because they found the presence of *KE* even when examining a group of patients who exhibited emotional processing deficit due to their ventromedial prefrontal lesion.

Later, Ngo et al. (2015) revaluated this interpretation, finding an increased amygdala activation (traditionally involved in emotion processing) associated with higher ascriptions of intentionality for the negative consequences; therefore they concluded that emotions play a key role in driving this kind of intentionality judgment, and this cerebral area was involved instead of the ventromedial one. Moreover, the authors found a relationship between judgments of *emotional reactions* after reading a series of Knobe-like scenarios and *intentionality ratings* provided in the Harm cases.

Finally Cokely and Feltz (2009) also found a positive correlation between the extraversion personality trait and the asymmetry width and concluded that, since the extraversion trait is related to emotional expressiveness and looser regulation of affective reactions, it could also be interpreted as evidence in favour of the affective bias hypothesis.

1.3.4 The deep self and true-self hypothesis

Another explanation was that introduced by studies of Sripada and Konrath (2011) who posited that asymmetric judgments are driven significantly by assessment of the Chairman's values, attitude and behavioural dispositions. When asking participants to rate the Chairman on a scale from "very anti-environment" to "very pro-environment", their results showed that participants are more inclined to classify an action as intentional if this action is concordant with the agent's attitudes, according to the *Deep Self Concordance Model* (Sripada, 2010) (e.g., that they are more inclined to regard the action of harming the environment as intentional if the agent is seen as anti-environmental). Similarly, studies of Hughes and Trafimow (2012, 2014) suggested that inferences about an actor's character and motives influence intentionality attributions about side effects, finding for instance that intentionality attributions were greater for good side effects when an actor was described positively and had positive motives.

Lastly, Newman, De Freitas & Knobe (2015) suggested that intentionality asymmetries could be more explained by beliefs about the agent's "true self" (a person's "core" or "essence"), who the authors distinguish with respect to agent's attitudes invocated by Sripada and Konrath (2011).

As beyond the scope of our project, we do not explore further models. For additional information see the review of Feltz (2007).

1.4 The role of individual differences of the person who judges on intentionality attribution and bias

Despite a large number of studies, research on the *KE* has mainly neglected the potential contribution of individual differences of the person who provides the intentionality judgments, with the exception of a few interesting studies, which outlined the pervasive and significant effect of individual characteristics.

(i) Nichols and Ulatowski (2007) suggested that the asymmetry may be the result of specific individual differences in interpreting the word 'intentionally'. Indeed the authors replicating the *KE*, noticed that a number of participants dissenting from the majority responses: a third responded "no" to harm and help, a third responded "yes" to both, and a third responded "yes" to harm and "no" to help. When asked to elucidate their decisions, participants' explanations concerned: the "lack of desire by the Chairman" from the participants who judged unintentional both side effect; whereas "the predictability, so the belief about the consequences of side effect" explained the decisions of the other two groups. According to the authors, these interpretations are reflective of two separate concepts of intentional action—in the first one, a knowledge about the consequences of the action is sufficient for a judgment of intentionality (*knowledge-based concept*), whereas in the second one also a desire/motive is needed to consider an action as intentional done (*motive-based concept*) (2007, p. 361).

(ii) This intuition was explored in more detail by Mele and Cushman (2008), who corrected the methodological shortcoming of Nichols and Ulatowski's study (2007) [2] , giving participants different scenarios where the protagonist (1) has a belief yet lacks a desire, or (2)

[2] who employed only the classic Knobe scenario, which is characterized by presence of belief about consequences but lacks desire, so did not allow to test if desire is sufficient or belief is necessary condition to act intentionally.

has a desire but lacks a belief. Like Nichols and Ulatowski, they found three patterns of responses—some who answer “yes” to both harm and help, some who answer “no” to both, and some who answer “yes” to harm but “no” to help. In scenarios where a person has a desire but lacks a belief, almost all participants retained that having a desire is sufficient for acting intentionally, whereas participants differ on whether belief is sufficient for acting intentionally. These results led authors to suggest the presence of “at least two concepts of intentional action”: one that consider belief as a sufficient condition (explain answers of “yes” to harm and help) and one that treats desire as a necessary condition (explains answers of “no” to harm and help). Moreover they speculated about the existence of a third concept which explains the asymmetric answers, according which people consider desire as a necessary condition for acting intentionally, except for morally bad actions; in such cases, belief becomes a sufficient condition.

(iii) Later, Cokely and Feltz (2009), as mentioned before, took into account again the role of individual differences in attributing intentionality among people, asserting that stable individual differences play a significant role in a variety of domains, such as the moral one (Feltz & Cokely, 2008a, 2008b; Feltz et al., 2009; Haidt, 2007). In this case, the authors considered the contribution of personality traits, due to their significant role in a broad range of decision-making processes (Funder, 1995, 1991; McCrae & Costa, 1990). In particular they examined the role of extravert personality, expecting a greater sensitivity to the asymmetry of judgments, due to the association of this trait with looser regulation of affective reactions and greater sensitivities to social dynamics. Consistent with their hypothesis, they found that people with an extravert personality show a greater asymmetry of intentionality judgments in the Knobe scenario, compared to introverts. Moreover, they found an association between the extraversion personality trait and the *belief-is-sufficient concept* whereas introverts tended towards a *belief-is-insufficient concept*. Those who held the *belief-is-sufficient concept* tended to judge that all side effects were more intentional whereas those who held a *belief-is-insufficient concept* tended to judge that all side effects were less intentional.

(iv) A further interesting evidence about the role of individual differences was provided by the study of Nduibuisi and Byrne (2013), who showed the influence of counterfactual reasoning on intentionality ascription. Counterfactuals are mental alternatives about how a past negative outcome could have been different, which are usually produced by individuals after the occurrence of a negative event (Byrne, 2016). The authors suggested that in the harmful situation of the *KE* exists a dilemma between the increase of profit and the side effect, whereas in the positive one there is no dilemma, that is the protagonist has a choice when faced with the harmful dilemma but not with the helpful one.

According to the authors, the dilemma of the harmful situation leads people thinking about alternatives the protagonist could have undertaken, so they perceive that he had a choice and they judge the side-effect to be intentional; in contrast, in the helpful situation they do not think about alternatives, so they perceive the protagonist having little choice and judge the side-effect to be unintentional.

Thinking about choices depends on the individual ability of individuals to generate mental alternatives, that is counterfactual reasoning (they imagine a counterfactual alternative of not pursuing the goal and no harmful side-effect). In this instance, the more one individual produces counterfactuals, the more alternatives choices are produced, and the more intentionality could be attributed (*counterfactuals*→ *choices*→ *intentionality*). Indeed, if participants are required to produce counterfactual thoughts prior to making their judgments of intentionality, the side-effect asymmetry is amplified, because they think of alternative choices the protagonist could have made (Byrne, 2012).

It is important to note that prior research (e.g., Kasimatis & Well, 1995) identified a different propensity, among people, to engage in one type of counterfactual compared to the other, as well as to be more or less likely to engage in counterfactuals reasoning at large (for instance individuals with high self-esteem engage more in downward counterfactuals compared to the upward ones).

These studies suggested us that the way to interpret the concept of "intentionality" can vary greatly among people due to their individual differences, therefore it could be simplistic assume the existence of a single explanatory mechanism of intentionality applying to all individuals. Nevertheless, the role of individual differences is mostly still under-investigated.

1.5 Aim of the present dissertation

The topic of the present dissertation was explored in four experiments which investigated certain individual differences that have a significant impact on intentionality attribution.

The first study explored, in more detail, the role of individual disposition to produce *counterfactuals* on intentionality judgments. Indeed, these mental alternatives may concern how an event could have been better (*upward counterfactuals*), as well as how an event could have been worse (*downward counterfactuals*) but previous studies did not distinguish between the two with respect to intentionality (see Nduibuisi & Byrne, 2013; Byrne, 2012). Therefore Study 1 investigated how the production of better (upward) and worse (downward) counterfactuals, compared to what actually happened, influences attributions of intentionality, as well as of responsibility and length of punishment (not yet been investigated as well). In this case, a given criminal event is given to be judged (since Byrne, 2012, Nduibuisi & Byrne, 2013 already investigated the Knobe scenario).

Then Study 2 explored another individual characteristic, closely associated to intentionality attribution, because it involves the understanding of others' mental states: the individual *Theory of Mind ability* of the person who judges. In fact the *KE* is characterized by a reduced attention to information about belief/intention of the Chairman compared to the attention paid on side effects, and Theory of Mind (ToM) is of particular interest because it represents the ability, ranging among people (Hughes et al., 2005), necessary to integrate belief information (Koster-Hale et al., 2013; Young et al., 2007). Its role has been widely studied in the domain of morality, where several studies showed that a more advanced understanding of the mental states of others allows to produce more mature moral judgments based on belief information (e.g., Fu et al., 2014). Despite its close connection, how individual differences on this capability are reflected in the *KE* and on intentionality ascription had not yet been investigated.

In order to further examining the Theory of Mind ability, in Study 3 testing individuals who present *autistic traits of personality,* we explored how intentionality judgments are affected when this ability is reduced. Indeed, according to recent studies (Gökçen et al., 2014, 2016), these individuals showed atypical theory of mind abilities, milder but comparable, to individuals affected by autism spectrum disorders who, by definition, are characterized by reduced theory of mind abilities.

Lastly, in Study 4, the role of individual differences in processing emotions on intentionality attribution was explored. Indeed, several studies supported the "Blame and Affective hypothesis", claiming for a role of emotions in influencing intentionality ascription and the *KE*, determining the higher judgments provided to the negative side effect (e.g. Malle and Nelson, 2003). However some contrasting results to this hypothesis were also obtained. Therefore, in the last study of the present dissertation, we tested individuals who present a*lexithymic trait of personality*, which involve difficulties of emotional processing, in order to provide further evidence about the role of emotions on intentionality attribution.

Chapter 2. Individual differences in intentionality attribution

Experiment 1: The role of downward and upward Counterfactuals on intentionality attribution

1. Introduction

As introduced before, counterfactual thinking has an important role because it permits to mentally create alternatives to reality in order to think how negative events might have turned out if only something had been different (Byrne, 2016). This type of reasoning is common in everyday imaginative thought and refer to spontaneous thinking about the possible alternatives of events that have already occurred, constructing hypothetical scenarios that could change those events. For example, if a man plays every week the same numbers in the lottery, and his numbers win the only week he forgot to play, he may spontaneously think "if only I'd played my numbers, I would have won".

Most studies have found that counterfactuals are linked to judgments about responsibility, showing that these judgments are more severe when changing a behaviour would lead to a better outcome, compared to when changing the same behaviour would not (e.g. Branscombe, Owen, Garstka, & Coleman, 1996; Mandel & Dhami, 2005; Nario-Redmond & Branscombe, 1996; Turley, Sanna, & Reiter, 1995; Wells & Gavanski, 1989). However, the comparison between better (upward) and worse (downward) alternatives with respect to what actually happened has not been investigated extensively on attributions of responsibility (but see Catellani & Bertolotti, 2014). Moreover, previous studies did not consider the effects of the type of counterfactual (upward versus downward) on judgments about other people's intentions and the consequent need for punishment.

Therefore, the present study sought to explore the association between upward and downward counterfactuals and the attribution of responsibility to the actor, the perception of his/ her intentionality and the ascription of punishment.

Indeed, recent evidence (Kasimatis & Wells, 2014) outlined that people differently engage in one type of counterfactual, compared to the other, according to their individual characteristics. For instance, different personality dispositions such as self-esteem or rumination, showed to influence this engagement, the former inducing more to downward reasoning, whereas the latter inducing more to the upward one.

The other aim of this study is to further understand the reasoning processes underlying counterfactual thinking examining the mediation role of the perception of predictability and the possible influence of the experimental conditions (upward versus downward) on the content of the counterfactuals, distinguishing between defendant/victim's behaviours and context factors.

1.1 Upward and downward counterfactuals

Research in counterfactual thinking has touched on several particular types of counterfactuals. In particular, Markman and collaborators (Markman, Gavanski, Sherman, & McMullen, 1993; McMullen, Markman, & Gavanski, 1995) used the term upward to describe a counterfactual concerned with how the outcome could have been better, and downward to refer to a worse outcome. These kinds of counterfactuals, deeply studied in risk decision-making, are spontaneously generated depending on the specific emotion provoked by a particular outcome (e.g. Habib et al., 2012). For example, when an individual obtains a negative outcome, such as a loss, he/she is more likely to feel regret and to produce upward counterfactuals, while when s/he escapes a danger it is more probable that s/he feels relief generating downward counterfactuals. The upward counterfactuals are the most common types of thoughts used and they may elicit intentions to perform success-facilitating behaviours, enhance task persistence and improve performance (Epstude & Roese, 2011; Markman, McMullen, & Elizaga, 2008; Roese & Olson, 1995; 1997). In this sense, upward counterfactuals' serve to reflect on internal and controllable elements, such as one's behaviour, provoking negative affects like regret and sadness, in order to potentially improve actions in the future, whereas downward counterfactuals primary function is to improve mood and feeling better focusing on external and uncontrollable events, such as catastrophic scenarios (e.g., Epstude & Roese, 2008; Markman & McMullen, 2003; Summerville & Roese, 2008).

Specifically, frustration for the outcome, typically elicited by upward counterfactuals, is considered a signal that a goal has not been attained and thereby increases improvement motivation, whereas outcome satisfaction, generally elicited by downward counterfactuals, enhances mood and diminishes the motivation to change one's behaviour (e.g., Zeelenberg, 1999). In conclusion, the functions of counterfactuals are several, such as preparing for future, explaining the past, eliciting emotions, such as regret and relief, and ascribing blame and fault, but the specific function of upward counterfactuals is to learn from mistakes, understanding the causes of bad outcomes and preventing them in the future, whereas the principal aim of downward counterfactuals is to justify past performances (Byrne, 2002; Markman & McMullen, 2003). Therefore, upward and downward counterfactuals are goal-driven thinking processes underlying different kinds of reasoning. Specifically, upward counterfactuals may be more linked to causal reasoning that occurs when people try to identify potential causes of bad outcomes (e.g., Tasso, 1999; Wells, Taylor, & Turtle, 1987). In this case, the comparison between reality and a counterfactual scenario generates causal inferences, for example "if I had not stumbled I would have not broken the leg, thus I broke my leg because I stumbled". Downward counterfactuals are usually more related to contrastive reasoning, that is, the comparison of a factual situation to an alternative one in order to justify bad outcomes (e.g., McGill & Klein, 1995). This type of reasoning is used when it is useful to consider the difference between the circumstances in which a certain event occurred (e.g. I was hurt to extinguishing a fire) and the circumstances in which the situation would have been worse (e.g. if I had not promptly acted, the fire would have destroyed my house). Counterfactual thoughts serve not only to assess one's own actions but also those of others. In the evaluation of a criminal event, for example, the judge may use his/her counterfactual thinking to evaluate a particular crime and to assign a punishment. He/she may focus particularly on internal factors, that is information about the defendant and/or the victim, including their counterfactual actions (i.e. what the defendant or the victim could have done but did not, or they could have not done but did), or external elements, that is contextual factors or the behaviour of other people besides the defendant or the victim.

1.2 Responsibility, predictability and counterfactuals

The relationship between counterfactual thinking and responsibility attribution, and consequent need for punishment, has been fully demonstrated in literature. In a classical experiment, Macrae, Milne, and Griffiths (1993) used a scenario where a woman contracted a disease from eating in a restaurant, that she regularly attends (routine case) or she had never attended (exceptional case), showing that, in the exceptional case, not only restaurant owners had to pay a greater refund, but they were even judged more guilty. Moreover, the greater the counterfactual thoughts are related to the defendant or the victim, the greater the responsibility is given, respectively, to the defendant or to the victim (e.g., Branscombe et al., 1996; Wiener et al., 1994). Similarly, Turley et al. (1995) conducted four studies on a crime of sexual abuse. Results showed that in the exceptional conditions (e.g. walking home via an unfamiliar new route), opposed to habitual ones (e.g. walking home via usual route), judges generated more counterfactual thoughts and, consequently, the judgments of responsibility and the degree of punishment for the offender increased when the exceptional behaviour was that of the defendant, whereas they decreased when the exceptional behaviour was that of the victim. Previous research has considered the effects of counterfactuals focused on actions rather than omissions, as well as the effects of counterfactuals focused on conforming rather than non-conforming actions/omissions on attribution of responsibility (for a review, see Catellani & Milesi, 2005). Recently, Catellani and Bertolotti (2014), considering the effects of counterfactual defences employed by politicians, showed that people exposed to downward rather than upward counterfactuals attribute less responsibility to the actor focused in the counterfactual. Generally, there is evidence that focusing on counterfactual thoughts, rather than factual, increases the perception of responsibility (e.g., Mandel & Dhami, 2005). However, there is also evidence that is not consistent with the idea that counterfactual thinking and responsibility attribution are associated (e.g., Mandel, 2003; Mandel & Lehman, 1996; Marques et al., 2014). Furthermore, a cognitive assessment of the perceived predictability of the negative outcome seems important for the attribution of responsibility.

Accordingly, a recent study showed that the more people think to what a defendant could have done but did not, the more they think that the negative outcome could have been foreseen and thus avoided; this, in turn, leads to heighten responsibility of the defendant (Catellani, Alberici, & Milesi, 2004). For example, in the case of a woman who is coming home through a usual route and gets injured in a car crash by a drunk man who invades her lane, people think that the man could have foreseen, and thus avoided, the car accident. In this case, people judge the man more responsible for the car crash compared to the case in which they know that the route, where the accident occurred, is rarely covered by the woman (Davis, Lehman, Wortman, Silver, & Thompson, 1995). In the first scenario, people generate more counterfactuals about the man (e.g., “if only he had not drunk before driving”), whereas in the second scenario, they generate more counterfactuals about the woman’s decisions (e.g., “if only she had taken her usual route”).

This example helps to better understand how the perceived predictability of a negative outcome, for the defendant, may play a mediating role in the relationship between counterfactual thoughts and responsibility attribution.

1.3 Intentionality and counterfactuals

Knobe (2003) showed that people provide intentionality judgments also for side effects, which by definition are not intentional, and secondly that such judgments are especially provided for negative side effects rather than positive ones. According to the literature, Nanay (2010) argued that it is easy to "slip from counterfactual dependence to causal attribution of responsibility and, therefore, of intentionality" showing that the attribution of intentionality is linked to the alternative scenarios that are elicited, or not elicited, in the mental representation of the events.

Consistently recent studies showed the crucial role of counterfactual thinking on judgments about other people's intentions (e.g., Knobe, 2010; Pellizzoni, Girotto, & Surian, 2010; Young & Phillips, 2011). For example, Knobe (2010) and Young and Phillips (2011) showed that counterfactuals mediate the relationship between moral and intentionality judgments: focusing on immoral actions, rather than moral ones, leads people to imagine more counterfactual alternatives to those specific actions and, as a result, to increase intentionality judgments; on the other hand, Pellizzoni et al. (2010) argued that counterfactual thinking itself, rather than the moral evaluation of a specific action, determines the intentionality judgments. Indeed, they showed that, in the Knobe scenario (2003), if the chairman does not know the side effects of his actions, the asymmetry between intentionality judgments (in negative and positive side effects conditions) is reduced, and it is eliminated when they are deemed to hold a false belief about the consequences of his actions. Finally, as stated before counterfactuals not only influence but also amplify judgments of intentionality (Ndubuisi & Byrne, 2013).

1.4 The present study

As mentioned above, previous studies showed that there is a link between the number of counterfactuals regarding defendant or victim and the degree of responsibility assigned to either of them (e.g., Mandel & Dhami, 2005; Nario-Redmond & Branscombe, 1996; Turley et al., 1995), but other studies have not found this relationship (e.g., Mandel, 2003; Marques et al., 2014). Moreover, limited attention has been dedicated to the effect of different kinds of counterfactuals on attribution of responsibility (e.g., Catellani & Milesi, 2005) and there is no research that has manipulated whether participants generate upward or downward counterfactuals before making judgments of responsibility. Moreover, in the literature, the type of counterfactual (downward versus upward) has not been considered together with the attribution of intentionality to other people. To address these gaps, we investigated the relationship between downward and upward counterfactuals and attribution of responsibility, intentionality and punishment. In line with recent studies (e.g. Segura, 2014), this study examined these relationships in a real and detailed scenario about a given criminal event.

Our first hypothesis is that focusing on downward or upward counterfactuals influences not only the attribution of responsibility but also perceived intentionality and recommended punishment. Specifically, according to previous studies (e.g. Catellani & Bertolotti, 2014), we hypothesized that when people think of worse scenarios, they may attribute less responsibility, intentionality and punishment to the defendant, than when they think of better scenarios. Moreover, we supposed that when people imagine worse scenarios they would be prone to think that the victim, as how the defendant, has a lower ability to predict the adverse outcomes and, consequently, a lower responsibility about what happened. Second, since previous studies (Catellani et al., 2004; Lagnado & Channon, 2008) showed a relationship between predictability and responsibility judgments, we expected that the perception of predictability mediates the relationship between counterfactual thinking, responsibility and intentionality judgments: for example, when participants are asked to imagine worse scenarios, compared to better ones, they may judge the actors less able to predict the outcome of their actions. These attributions about predictability, in turn, could influence judgments about responsibility, intentionality and need for punishment.

Finally, in order to examine the reasoning processes involved in counterfactual thinking more closely, we assessed whether the experimental condition (upward versus downward) influences the content of counterfactuals, distinguishing between external factors, such as situational factors or the behaviour of others characters, and internal factors, that is, the behaviour of the defendant and/or of the victim. We expected that when people imagine worse scenarios they focus more on external conditions, while when people imagine better scenarios they would produce more counterfactual thoughts about victim's and defendant's behaviours. This hypothesis is based on the characteristics of counterfactuals studied in literature, that is when people focus on downward counterfactuals they usually mutate external and uncontrollable events, whereas when they focus on upward counterfactuals they change internal and controllable elements in hypothetical scenarios (e.g. Epstude & Roese, 2008; Markman & McMullen, 2003; Summerville & Roese, 2008).

2. Method

2.1 Participants

Ninety-three individuals (42% men) were recruited for the study. The average age of the sample was 36.38 years (SD = 12.75), ranging from 20 to 60 years. All participants provided informed consent prior to the study and their participation was voluntary. Demographic variables concerned education (23.7% secondary/high school, 76.3% bachelor's degree or graduate degree) and experience in judicial topics (45.2% of the sample was inexperienced, 54.8% studied judicial topics or worked in the field of law). The study was approved by the Ethics Committee of the University of Bologna.

2.2. Materials and procedure

Participants completed a booklet, created for the study. After completing the demographics questions, participants were presented a two-page report of an assault case, based on a true case and used in Catellani and Milesi (2001). Two men, Stefano and Mauro, were driving along different routes; they noticed that the traffic was heavy and that they were both late. Therefore, they decided to change their route for a less busy one. They came to a crossroads after which they both had to take the same road. Mauro entered the road without stopping at the give-way signal, cutting in front of Stefano, who had to brake suddenly. At this point, both of them got angry: Stefano tried to pass Mauro, who suddenly braked, then Stefano bumped into Mauro's car. The men went out their cars and started arguing more and more sharply. Suddenly, Mauro, who is a plumber, flew into a rage and grabbed Stefano's arm, taking a monkey wrench out of his pocket. Stefano raised his arm to protect himself and after having been hit by the wrench collapsed. Mauro ran to his car and drove away. Passers helped Stefano who was hospitalized for 15 days. Afterwards, Stefano charged Mauro with assault. Mauro admitted to having hit Stefano but in self-defence.

Participants were randomly assigned to the two experimental conditions, upward (N = 45) and downward (N = 48). They were asked to "write as many possible alternatives if the story had gone better (upward condition) or worse (downward condition) comparing to the actual situation". Successively, participants were explained that for the Italian Penal Code Mauro had been charged with two different crimes[3]: the assault crime, that refers to the aggression against Stefano, and the personal injury crime, that results from the physical injury caused by Mauro to Stefano, who consequently got hospitalized. Participants were asked to imagine being the judge of the trial and to give judgments about the two crimes, answering, on a scale

[3] For the Italian Penal Code the two crimes are actually different because they offend specific goods and they are not necessarily linked. Specifically an assault might not produce personal injury; consequently, in the case in which the defendant causes personal injuries to the victim both the two crimes have to be judge (as in the present scenario).

of 0 (not at all) to 100 (absolutely), a few questions regarding: the predictability of outcomes (i.e. assault crime: “How much predictable was the aggression of Mauro for Stefano?”; personal injury crime: “How much predictable for Mauro were the health consequences of Stefano?”), the responsibility (assault crime: “How much responsible was Mauro for his aggression?”; personal injury crime: “How much responsible was Mauro for the health consequences of Stefano?”), the intentionality (assault crime: “Did Mauro intentionally assault Stefano at that time?”; personal injury crime: “Did Mauro intentionally provoke physical injury to Stefano?”). Moreover, participants were asked to assign the length of punishment for the assault crime (“The penalty for this crime is from 0 to 6 months of imprisonment; which punishment would you assign to Mauro?”) and for the personal injury crime (“The penalty for this crime is from 0 to 3 years of imprisonment; which punishment would you assign to Mauro?”). Successively, for the data analyses, these two temporal responses were transformed on a scale from 0 to 100. Finally, participants’ accounts were analysed in order to distinguish the counterfactuals depending on the content. A distinction was made between counterfactuals focused on external factors, such as situational factors (e.g. Participant # 3: “if there had not been traffic”) or the behaviour of other characters (e.g. Participant # 55: “if someone had intervened to quiet the quarrel”), or internal factors, such as the behaviour of the victim (e.g. Participant # 22: “if Stefano had not tried to pass Mauro”) and/or of the defendant (e.g. Participant # 87: “if Mauro had not hit Stefano”). Two independent judges, unaware of hypothesis, carried out the coding, with an agreement of 97%. Any discrepancies in coding were resolved through discussion.

3. Results

3.1 Manipulation check and descriptives

We checked that participants effectively generated counterfactuals as instructed, depending on what condition they were in. Every participant assigned to the upward condition only produced upward counterfactuals and every participant assigned to the downward condition only produced downward counterfactuals. Overall, participants generated a mean number of 5.13 counterfactuals (SD = 3.19).

Means, standard deviations and intercorrelation of all the measures for the total sample are reported in Table 1.

	Means (SD)	Intercorrelations						
		1	2	3	4	5	6	7
Assault crime								
1. Victim's predictability	32.21 (21.61)	-	-	-	-	-	-	-
2. Defendant's responsibility	87.32 (13.22)	-.02	-	-	-	-	-	-
3. Intentionality	68.55 (25.55)	.04	.33**	-	-	-	-	-
4. Length of punishment	74.00 (26.05)	-.19	.13	.15	-	-	-	-
Personal injury crime								
5. Defendant's predictability	70.63 (25.32)	-.04	.19	.32**	.24*	-	-	-
6. Defendant's responsibility	90.26 (14.04)	-.13	.67**	.33**	.16	.29*	-	-
7. Intentionality	53.41 (29.25)	.03	.17	.33**	.38**	.42**	.27*	-
8. Length of punishment	46.97 (29.69)	-.02	.04	.17	.63**	.21*	.11	.38**

Table 1. Means, standard deviations and inter-correlations of all the dependent variables (**p*<.05: ***p*<.001)

In order to assess the associations between demographic variables (i.e. gender, education and experience in judicial field) and dependent variables (i.e. perceived predictability, responsibility, intentionality and length of punishment for the defendant), we performed a multivariate between-subjects ANOVA for each crime. We performed two MANOVAs because the dependent variables are conceptually related and there are some significant intercorrelations between them (see Table 1). No gender differences were reported in dependent variables (ps from .62 to .29), except for the defendant's responsibility in the assault crime [F(1, 92) = 9.39, p < .01, η2 = .17]. In particular, males (M = 90.81, SD = 2.53) had higher ratings of perceived responsibility than female (M= 77.72, SD = 3.21). Education variable was not associated with dependent variables (ps from .98 to .09), such as experience in judicial field (ps from .98 to .20). Moreover, considering the high variability of the age of the present sample, we performed a regression analysis for each dependent variable to evaluate whether age predict them. The results showed that age did not predict any of the dependent variables (ps from .85 to .19).

3.2 Effects of upward or downward counterfactuals

For the crime of assault, we performed a multivariate between-subjects ANCOVA considering the experimental condition (upward versus downward counterfactuals) as independent variable, gender as covariate since it was found to predict responsibility in the previous analysis, and judgments about perceived predictability, responsibility, intentionality and length of punishment as dependent variables. Significant main effects of counterfactuals emerged for perceived responsibility [F(2, 91) = 4.43, p < .01, η2 = .05], perceived intentionality [F(2, 91) = 19.14, p < .001, η2 = .17] and for the length of the punishment [F(2, 91) = 3.42, p < .05, η2 = .04], whereas there were no significant differences between the experimental conditions on the perception of the perceived predictability [F(2, 91) = 1.11, p = .29]. For the crime of personal injury, we performed another multivariate between-subjects ANOVA considering the experimental condition (upward versus downward counterfactuals) as independent variable and judgments about perceived predictability, responsibility, intentionality and length of punishment as dependent variables. Significant main effects of counterfactuals emerged for defendant's predictability [F(1, 92) = 4.25, p < .05, η2 = .05], intentionality judgment [F(1, 92) = 7.42, p < .01, η2 = .08] and the length of the punishment [F(1, 92) = 4.59; η2 = .05]. No significant main effect emerged for perceived defendant's responsibility [F(1, 92) = 2.69, p = .10]. Means and standard deviations are shown in Table 2.

	Upward	Downward
Assault crime		
1. Victim's predictability	34.57 (24.23)	30.00 (18.82)
2. Defendant's responsibility	90.22 (10.75)	84.48 (14.19)
3. Intentionality	79.80 (20.38)	58.51 (25.83)
4. Length of punishment	80.00 (30.03)	68.33 (26.05)
Personal injury crime		
5. Defendant's predictability	76.35 (25.72)	65.27 (23.97)

6. Defendant’s responsibility	92.49 (12.78)	88.18 (14.96)
7. Intentionality	61.62 (28.83)	45.55 (27.72)
8. Length of punishment	52.33 (24.16)	44.55 (23.88)

Table 2. Means, standard deviations of dependent variables.

3.3 Mediation effect of predictability

To test for mediation effect, we used the approach laid out by Baron and Kenny (1986) that requires three regression equations. The first tests for a significant relationship between the independent variable (i.e. upward counterfactuals = 1; downward counterfactuals = −1) and the mediator (i.e. perceived predictability for the defendant). The second looks at the relationship between the mediator and the outcome variable (i.e. intentionality judgment). If both of these equations are significant, a third regression equation is computed in which the independent and the mediator variables are included as predictors of the outcome variables. Mediation would be either full or partial. Full mediation can be inferred if the regression coefficient for the predictor is not significant in the third regression equation, while partial mediation can be inferred if the value drops, but remains significant. Significance of mediation is evaluated using the Sobel test (Sobel, 1982). We only considered the scores on personal injury crime because of the lack of effect of condition on perceived predictability in the assault crime. In the first test for mediation, the regression analysis revealed a relationship between counterfactuals and perceived predictability for the defendant [$F(1, 92) = 4.62$, $p < .05$], accounting for 4% of the variance (see Figure 2).

In the second test for mediation, the regression analysis showed a predictive effect of perceived predictability on intentionality judgment (β = .42, p < .001), [F(1, 92) = 19.29, p < .001], accounting for 16% of the variance. In the third test for mediation, we examined whether the relationship between counterfactuals and intentionality judgment was attenuated when the mediator, that is predictability judgment, was included in the regression model. The analysis showed that intentionality was associated with both counterfactuals and perceived predictability [F(2, 91) = 11.96, p < .001]. The type of counterfactual accounted for 7% of the variance in the score of intentionality and predictability explained another 12% (see Figure 2). Calculation using the Sobel test indicated that perceived defendant's predictability (Z = 5.56, p < .001) partially mediated the relationship between counterfactuals and intentionality judgments.

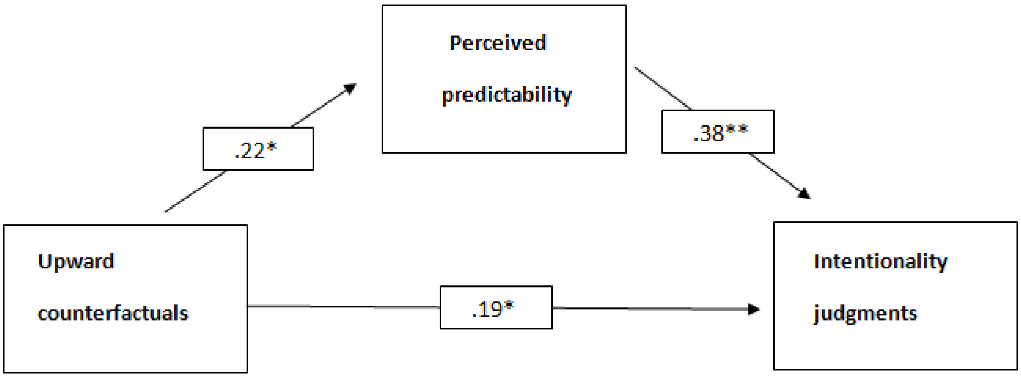

Figure 2. Mediation model about the personal injury crime (*p < .05; **p < .001).

3.4 Content of counterfactuals

20.42% of counterfactuals referred to external factors (i.e. situational factors or the behaviour of other characters), and the remaining 79.58% focused on internal factors (i.e. behaviour of the victim and/or of the defendant). A multivariate between-subjects ANOVA showed that demographic variables (i.e. gender, education and experience in judicial field) were not associated with the content of counterfactuals (ps ranging from .09 to .91). Moreover, we performed two regression analyses to evaluate whether age predicts the content of counterfactuals (i.e. external factors or internal ones). No statistically significant effects were found (ps ranging from .36 to .97). In order to assess whether the experimental conditions (i.e. upward versus downward) were associated with the content of counterfactuals (i.e. external or internal elements), we performed a multivariate between-subjects ANOVA. Results showed a significant main effect of the condition on both internal [$F(1, 92) = 26.44$, $p < .001$, $\eta2 = .23$] and external contents [$F(1, 92) = 29.61$, $p < .001$, $\eta2 = .25$]. Specifically, counterfactuals were more focused on the internal factors in the upward condition ($M = 6.71$, $SD = 3.00$), rather than in the downward one ($M = 3.52$, $SD = 2.81$), whereas counterfactuals were significantly more focused on external factors in the downward condition ($M = 3.24$, $SD = 2.85$), rather than in the upward condition ($M = 0.75$, $SD = 1.46$).

4. Discussion

The first aim of the study was to explore the association between upward and downward counterfactuals and participants' self-reported perceived responsibility, intentionality and need for punishment of an individual. The results showed close relationships between the type of counterfactuals and these attributions, confirming that it determines both responsibility and intentionality judgments, and the consequent assignment of punishment (e.g. Macrae et al., 1993; Pellizzoni et al., 2010; Turley et al., 1995), offering some elements of interest for the study of the decision-making in the legal field.

As hypothesized, the production of downward counterfactuals leads people to judge the criminal event less intentional, the defendant less responsible and, therefore, to give him a less severe punishment. In this sense, imagining worse outcomes for the victim, as his death or some additional injury, implies that what did occur was not as bad as it might have been. Consequently, it seems that the perceived severity of the original event is reduced through the comparison with hypothetical more dramatic events, which become an important information on which people rely to provide their judgments. People in fact do not give absolute judgments but they base their evaluations on what could have happened, that become an anchor for the subsequent judgments (Tversky & Kahneman, 1982). As mentioned above in the Introduction section, some authors distinguished between contrastive and causal reasoning underlying downward and upward counterfactuals respectively (e.g. Tasso, 1999). This distinction is useful to better understand how events are interpreted and how judgments about others arise. In the case of downward counterfactuals, people could reason following a contrastive principle comparing the circumstances in which an event happened with the circumstances in which it has not happened (e.g. McGill & Klein, 1995). For example, to assess the severity of the consequences of a car accident, an observer would compare the circumstances in which the event occurred, such as the fact that the driver had the seat belt, with other counterfactual circumstances, such as the hypothetical scenario in which, without the seat belt, the driver could get a much worse injury. In this sense, a factual outcome may appear better when a hypothetical less desirable outcome becomes salient (e.g. Markman & McMullen, 2003; Medvec & Savitsky, 1997; Mellers, Schwartz, Ho, & Ritov, 1997; Roese, 1994). On the other hand, the production of upward counterfactuals leads people to perceive the defendant more responsible, to evaluate his behaviour more intentional and, consequently, to give him a more severe punishment. Such as the previous one, this reasoning is related to the comparison with counterfactual alternatives, which leads people who judge to attribute negative intentions to the defendant and to perceive him more responsible. However, in this case, an observer makes another kind of reasoning besides the contrastive one, that is thinking there is a causal link between the actions taken by the defendant before the event and the negative outcomes (i.e. aggression and hospitalization). This causal reasoning occurs when people identify a possible relationship between a potential cause and a specific outcome through the production of counterfactuals (e.g. Tasso, 1999). In the previous example, to assess the causes of the car accident, an observer might compare the actual circumstances, such as a distraction of the driver, with better counterfactual circumstances, such as the scenario in which the driver would have avoided the car accident if he had been more careful. Causal inferences occur because a counterfactual thought may dramatize and emphasize this link (Roese, 1997). As a matter of facts, focusing on the antecedent behaviour of the defendant brings an observer to feel more sense of injustice about the negative outcomes and to judge the defendant more responsible and to assign him more severe punishment. In literature, it is already shown that counterfactuals amplify responsibility attributions. For example, Nario-Redmond and Branscombe (1996) showed to participants an identical rape event, which could have been, respectively, worse for the victim or worse for the defendant by imagining how either the victim (e.g. if only she rebelled, she could die or have some additional injury) or the defendant (e.g. if only he was less careful, a relative of the victim could see him, and the rape could have been vindicated) could have behaved differently. According to the results, when participants imagine an alternative outcome, that is even worse for the victim than the original rape event, the perceived injustice and the severity of the rape is reduced, rendering the defendant to be perceived as less responsible. Conversely, when participants imagine an alternative outcome, that is worse for the defendant than the original rape event, the perception of injustice increases and the attribution of legal responsibility serves to rectify this sense of injustice. Our data showed that the properties of counterfactual thoughts to activate contrast or causal implications not only directly activate responsibility attributions but also the perceived corresponding behavioural intentions and, consequently, the need for punishment. These findings are consistent with studies that assign a functional meaning to the counterfactual thinking: downward counterfactuals have the function of improving the observer's mood, relieving the anxiety generated by a negative event, whereas upward counterfactuals, rising regret for what happened, are used to help the construction of future scenarios where different behaviours would permit to avoid negative outcomes (e.g. Epstude & Roese, 2008). The results of the present study not only showed a relationship among counterfactuals and judgments about responsibility, intentionality and severity of the punishment but also offered some interpretations for these associations. In particular, we have considered two factors: the observer's perception of predictability for the defendant and the focalization on internal versus external circumstances. As regarding the first factor, this study showed that participants, who observe a criminal event, judge the outcome of the event less intentional when they have to think about bad scenarios also because they perceive that the defendant could less predict the negative outcome. On the other hand, participants judge the outcome more intentional when they think about better scenarios, attributing to the defendant the capability to foresee that the event could end badly. Some authors had already shown that the predictability mediates the relationship between counterfactuals and responsibility judgments (e.g. Catellani et al., 2004), but in previous studies there was not a distinction between upward and downward counterfactuals and participants were not external observers but they were involved in the event, interpreting the role of the defendant or of the victim. Thus, imagining worse scenarios lead people to judge the negative consequences less foreseeable for the defendant, and, then, to attribute him less intentions for such outcomes. Conversely, when people think about better scenarios, they focus on what the defendant could have done and they are more prone to think that negative results could have been avoided thus attributing greater intentionality to the defendant. Future studies could investigate this issue manipulating the perceived predictability in order to test its influence on intentionality judgments. Similarly, as regard the counterfactual mutation of internal versus external circumstances, our data showed that downward counterfactuals are linked to a greater focus on external circumstances, that could have even produced negative outcomes, while upward counterfactuals lead an observer to more focus on those behaviours of the defendant (and/or the victim) which could have improved the situation. Therefore, it seems that the content of a downward counterfactual drives the observer to focalize on external circumstances and to perceive the defendant as less able to foresee the negative consequences of his behaviour; this reasoning, that could be due to a contrast effect, leads the judge to diminish intentionality judgments, giving a less severe punishment to the defendant. On the contrary, an upward counterfactual, via a causal effect, pushes the observer to mutate the behaviour of the defendant, which is perceived as more able to foresee the negative consequences of his behaviour and, in turn, to give him a more severe punishment.

This research showed that when people have to judge others' responsibility and intentionality, the focalization on characters' actions, or external events, may influence these perceptions and the consequent need of punishment.

As regards the lack of significant effect of the experimental condition (upward and downward counterfactuals) on responsibility of the defendant in the crime of personal injury, we must emphasize that people distinguish the judgments of intentionality and responsibility, even if they are conceptually related, thus counterfactuals could affect one of the two judgments and not the other. For example, a driver who goes out of control while drunk and injures a family is judged to be legally responsible, but not to have harmed the family intentionally (Knobe, 2003), while if a person is attacked in the darkness and he reacts injuring the aggressor is judged to have inflicted the pain intentionally, for legitimate defence, but is not legally responsible for doing so (Sverdlik, 2004). Moreover, there is a fundamental difference between the two crimes proposed in the present study, that may affect the judgments of responsibility: in the crime of assault, the discussion between the two actors is due to the behaviour of both and, then, the observer can judge a shared responsibility to the crime; instead in the crime of personal injury, the consequences of the actions of the defendant (i.e. aggression) are quite severe (i.e. hospitalization), and for this reason, the defendant has clearly passed a certain threshold of severity, becoming totally responsible for those consequences. Therefore, with respect to the second crime, the judgment is believed more certain compared to the first crime and it is independent from the generation of counterfactuals.

In conclusion, although there are some limitations to this research, such as the issue that the findings are based on just one scenario, the present study establishes a relationship between the type of counterfactuals and intentionality judgments and suggests that this relationship is partially mediated by the perception of predictability and is related to the focalization on specific elements, such as external circumstances or internal, such as the behaviours of the defendant and/or the victim. The association between upward and downward counterfactuals and intentionality judgments could be a worthy addiction to the literature because of theoretical and practical implications. From a theoretical point of view, this study may give some suggestions about the presence (or not) of a relationship between counterfactual thinking and responsibility attribution discussed in the literature (e.g. Mandel, 2003; Mandel & Dhami, 2005; Marques et al., 2014). Specifically, it seems that not the number, but the type of counterfactual thoughts (downward or upward) is relevant in the attribution of responsibility. Moreover, the present study has also implications for a better understanding of the reasoning processes, as the perception of predictability of a specific outcome, that lead an observer to the attribution of intentionality to a specific behaviour. From a practical point of view, there are many potential applications of these findings. For example, this study could be helpful in the legal field for attorneys because it provides some guidelines on how they could ask questions to victims and defendants, focusing on worse scenarios than what really happened and on external circumstances rather than the defendant’s behaviour, in order to build a better defence in criminal cases. On the other hand, jurors and prosecutors could identify some factors that could influence their perceptions of experts, eyewitnesses and other legal actors, becoming more aware of their reasoning and providing more objective judgments about criminal responsibility, civil negligence and liability for punitive damages.

It is important to note that, in the present study, participants are randomly split up in the two experimental conditions (upward and downward counterfactuals), instead of looking for individual dispositions, in order to be able to study the influence of both on intentionality and responsibility judgments.

However, after have identified more situational factors that influence counterfactual thinking (e.g., exceptional events, the first event in a causal sequence, controllable events, actions), more recently research has constantly showed that the way in which people mentally simulate alternatives can be significantly different, even if they have experienced similar outcomes (e.g., Kahneman & Miller, 1986; Medvec et al., 1995; Roese, 1997): for instance, Sanna (1998) found that when optimists think counterfactually they tend to generate more repairing downward counterfactuals compared to pessimists; coherently Kasimatis and Wells (1995) showed that both optimists and those high in self-esteem are less likely to generate upward counterfactuals. Essentially, viewing events in a positive manner seems to be related to more infrequent use of upward counterfactuals. Furthermore, Bacon et al. (2013) examined the association between individual differences in fantasy, finding that higher levels of fantasy proneness are correlated with higher levels of spontaneous counterfactual thinking; whereas Murphy (2005) examined the influence of Digman (1990) five-factors of personality, finding that higher levels of "neuroticism" are associated with the use of upward counterfactuals and "openness to experience" to the downward ones.

These studies demonstrated that individual differences have an influence over both the activation and type of counterfactual thoughts undertaken. Therefore future research should consider more how individual characteristics, such as personality traits or cognitive styles, lead an observer to generate each of the two types of counterfactuals and how this is, in turn, reflected on responsibility and intentionality judgments.

Experiment 2: The role of individual differences in Theory of Mind ability on intentionality attribution

1. Introduction

As stated in the introduction section, the *KE* represents a bias of intentionality attribution, since the side effect is considered intentional even if, by definition, it would not be. Indeed, in the instance described, the Chairman has no intention of damaging the environment, whereas his intention concerns the program to increase the company's profits. However participants' judgments are mostly "catch" by outcome information, rather than on that concerning protagonist's beliefs.

Despite a large number of researches focused on explaining such effect (see the general introduction), research has mainly neglected potential contributions of individual differences, with the exception of few studies (Cushman & Mele, 2007; Nichols & Ulatoski, 2007; Byrne, 2012; Cokely & Feltz, 2009) which underlined this significant effect and called to consider more the way these differences affect the focus of individuals' attention while analyzing social situations.

In particular, the ability necessary to integrate belief information is defined *Theory of Mind* (ToM; Koster-Hale et al., 2013; Young et al., 2007). It represents the ability to attribute mental states to oneself or to other people (Wimmer & Perner, 1983) and ranges among people (Hughes et al., 2005). It is composed by two components: Cognitive ToM refers to the ability to understand the beliefs of others and affective ToM refers to the ability to understand what others' are feeling[4] (Shamay-Tsoory et al., 2005). Moreover, in a recent model, Shamay-Tsoory et al. (2010) stated that cognitive ToM is a prerequisite for affective ToM, which also requires empathy skill, intended as the ability to share others' emotional states (Singer,

[4] Importantly, affective ToM should not be confused with emotional empathy, otherwise termed affective resonance or the ability to feel what another is feeling. Moreover, in the literature, there is a quite variable terminology, so that Theory of mind (ToM) is variably referred as cognitive empathy or mentalizing (Gillespie et al., 2017).

Critchley, & Preuschoff, 2009). Therefore, successful affective ToM ability requires the integration of cognitive ToM and empathy. Empathy ability plays a critical role in human interpersonal engagement and social interaction, since it permits individuals to resonate with others' emotions, and the lack or decrease of this ability is typical in several serious pathologies such as narcissistic personality disorder or psychopathy (Bird & Viding, 2014; Decety & Moriguchi, 2007).

The role of ToM in integrating belief information has been widely studied in the domain of morality. Several studies showed that from childhood a more advanced understanding of the mental states of others is necessary to produce more mature moral judgments. A clear example is provided by a study of Fu, Xiao, Killen, & Lee (2014) who administered a ToM task to 79 children, where the protagonist commits an accidental transgression and asked them for moral judgments about him. They found that children with higher ToM made better moral judgments, understanding the accidental nature of the transgressor's action, with respect to children with lower ToM. Instead, when moral scenarios present conflicting information about the outcome of an action and the intention of the actor, much younger children's moral judgments and justifications are determined by the action's outcome rather than the actor's intention (Zelazo, Helwig, & Lau, 1996). With the development of ToM ability, they become progressively able to integrate belief information in moral judgment, reducing moral condemnation of accidents, since a more robust mental representation of others' beliefs helps them to down-regulate the emotional arousal from salient harm outcomes (Yuill & Perner, 1998).

Moreover, recent neuroimaging studies also showed an association between a greater activation in the right temporoparietal junction, as well as a greater local grey matter volume in the left anterior superior temporal sulcus, both regions part of the ToM network, with greater consideration of the actor's intentions, and consequently with a higher forgiveness towards accidental harms in moral judgments (Koster-Hale et al., 2013; Patil et al., 2017).

1.2 The present study

Given the closely relationship between ToM ability and the capability to detect the accidental nature of an action in moral judgments tasks, the aim of the present study was to empirically evaluate whether individual differences in cognitive and affective ToM abilities, controlling for empathy ability, could also predict the judgments of intentionality in the Knobe scenario, allowing to focus more on intention's information rather than on the outcome one.

The first hypothesis was that a greater cognitive and affective ToM ability would permit a more accurate evaluation of the protagonist's thoughts and motivations. In particular, we expected a lowering of the intentionality attributed to the negative side effect and consequently a decrease of the asymmetry of intentionality judgments about side effects, made by people with higher cognitive and affective ToM abilities. In line with the hypothesis participants with a higher ToM would pay more attention to the information concerning the intention of character compared to that on the side effect.

Moreover, given the differences detected in literature between cognitive and affective ToM abilities, we want to investigate the role of each one, on this type of judgment. Furthermore, as previous studies (Shamay-Tsoory et al., 2010) suggested that empathy ability represents a prerequisite for the development of affective ToM, we measured the individual empathic ability of participants, controlling for its possible effect on the attribution of intentionality.

Finally, in order to examine the reasoning processes involved in intentionality ascription more closely, understanding to which information participants relied on, we assessed the explanations provided by participants to their judgments, distinguishing between judgments focused on protagonist's mental states (such as intention or motive) or on resulting side effect. Therefore the second hypothesis predicted that people with a greater ability to reason on mental states (higher cognitive and affective ToM abilities) would focus more on protagonist's mental states, lowering intentionality judgments about side effect, while people poorer in reasoning on mental states (lower cognitive and affective ToM) would focus more on the side effect, increasing intentionality judgments for it.

2. Method

2.1 Participants

To determine the sample size, we performed power calculation using GPower 3.1 (Faul et al., 2007), which defines a sample size of eight-five participants necessary to perform regression analysis with three predictors of interest: Cognitive and affective ToM and empathy abilities (effect size f2=.15; α=.05; power =.85).

Eighty-eight volunteers were recruited at university campus and city cultural associations through notices on social networks and on bulletin boards. Exclusion criteria were a history of neurological or psychiatric disorders. Two participants were excluded because they met the exclusion criteria. The final sample was set at eight-six participants (40 women, 46 men, students and workers, Mean age =42.07; SD=21.06, age range: 20-60 years).

All participants signed a written consent form before the study began. The study was approved by the local Ethics Committee of the University of Bologna.

2.2 Materials and Procedure

2.2.1 Questionnaires

Since we are interested in detecting different degree of ToM ability, the *Reading the Mind in the Eyes Test* (RMET, Baron-Cohen et al., 2001) and the *Short Story Task* (SST, Dodell-Feder et al., 2013) are employed. Both of them are, indeed, considered among the few reliable measures able to detect individual ToM differences in healthy individuals (Turner &

Felisberti, 2017; Dodell-Feder et al., 2013[5]; Vellante et al., 2013[6]). On the other hand, the most classical ToM measures (e.g. False-Belief task- Corcoran et al., 1997) have been used successfully to distinguish clinical populations from healthy participants but are showed to be mostly insensitive to more subtle ToM deficits. Reading the Mind in the Eyes Test was used to measure ToM affective ability and Short Story Task to measure the cognitive one.

Reading the Mind in the Eyes Test. In the *Reading the Mind in the Eyes Test* (Baron-Cohen et al., 2001) participants saw 36 pairs of eyes and for each one had to judge which of four adjectives best described the mental state being expressed through the eyes (for example "jealous, fearful, arrogant, odious"). The score ranges from 0 to 36. Photographs are displayed centrally and the four adjectives (one correct adjective and three distractors) are placed in the four corners of the paper sheet. The 36 experimental trials are preceded by a single practice trial.

Short Story Task. In the *Short Story Task* (Dodell-Feder et al., 2013) participants read an ambiguous short story about the relationship between two persons and have to assess their mental states. The task is composed of three parts:

(i) Eight questions assess the understanding of characters' explicit first-order (i.e., inferring the character's beliefs; *Why does Marjorie reply "Oh Nick, please cut it out! Please, please don't be that way!"?*) and second-order mental states (i.e., inferring what one character thinks about another character's beliefs or actions; *Why is Nick afraid to look at Marjorie?*). For

[5] SST-*Using Fiction to Assess Mental State Understanding: A New Task for Assessing Theory of Mind in Adults:* Inter-rater reliability for the mental state reasoning score ICC=.98; and for the comprehension score ICC = .90. Inter-rater agreement on the presence versus absence of a spontaneous mental state inference kappa=.86.

[6] RMET- *The ''Reading the Mind in the Eyes'' test: Systematic review of psychometric properties and a validation study in Italy*: Cronbach's alpha= .605. Test-retest reliability for the Eyes test, as measured by intraclass correlation coefficient = .833 (95% confidence interval=.745 to .902).

each question, the “*explicit mental state assessment”* score ranges from 0 (inaccurate) to 2 (full understanding). The maximum score is 16.

(ii) Five questions probed reader comprehension of factual story events (non-mental content). For each question, the “*comprehension”* score ranges from 0 (inaccurate) to 2 (full understanding), for a maximum score of 10.

(iii) One question assessed reader comprehension about spontaneous characters’ mental states (presence versus absence of a mental state inference produced by participants in the question: ‘In just a few sentences, how would you summarize the story?’). In this case, the “*spontaneous mental state inference*” score ranges between 0 to 1 (presence versus absence of a mental state inference).

The total score of Short Story task was obtained by the sum of *explicit mental state assessment* (0-16), *comprehension* answers (0-10) and the *spontaneous mental state inference* (0-1). The maximum score is 27. Questions are open-ended and the answers provided were analyzed, according to Dodell-Feder et al.’ paper guidelines (2013), by two independent judges, unaware of hypotheses, who carried out the coding with an agreement of 96%. Any discrepancies in coding were resolved through discussion.

Interpersonal Reactivity Index. The level of Empathy abilities of participants was measured by the Interpersonal Reactivity Index (IRI, Davis, 1980) which is composed by four scales: Perspective Taking (PT) measures the tendency to spontaneously adopt the psychological point of view of others in everyday life; Empathic Concern (EC) assesses the tendency to experience feelings of sympathy and compassion for unfortunate others; Personal Distress (PD) measures the tendency to experience discomfort in response to others’ extreme distress and Fantasy (FS) measures the tendency to imaginatively transpose oneself into fictional situations such as books and movies. Participants have to judge to what extent the items describe them, on a five-point scale ranging from 1 (“Does not describe me well”) to 5

("Describes me very well"). The subscale scores were subsequently added together to obtain a total score. The total score ranges from 0 to 135.

2.2.2 Scenario and explanations provided about judgments

Participants were presented with the classic Knobe scenario, reported in the table below (Table 3).

The English version was translated into Italian by an Italian native speaker, proficient in English. Then, the translated dilemmas were presented to an English native speaker proficient in Italian for the back-translation to English (Brislin, 1970).

Participants are requested to determine the level of intentionality of the side effect using a scale from 0 (unintentional) to 100 (completely intentional). In the statistical analyses, the intentionality scores provided for each side effect (negative and positive), and the difference between them, to obtain a measure of the asymmetry between them (asymmetry of intentionality judgments), were considered as dependent variables.

After each judgment, participants were asked to provide a brief explanation about their decision: "*Why do you judge the negative side effect as intentional or unintentional*?". A distinction was made between answers focused on the character's intention (e.g. "he wants to increase profits, he wants to be rich, he doesn't care about the side effect") or on the side effect (e.g. "the environment will be damage, he knows the environment will be damaged").

Two independent judges, unaware of the hypothesis, carried out the coding, with an agreement of 97%. Any discrepancies in coding were resolved through discussion.

2.2.3 Procedure

All participants were individually tested by a single researcher in a quiet room for about 45 minutes. Instructions were provided at the beginning of each test. Firstly, they are required to complete a demographic questionnaire about their age, gender and education level, then they completed the ToM tasks, answered to the two versions (negative and positive) of the classic scenario (see Table 3) and provide a brief explanation about their decision.

We followed a within-subjects design, in order to exclude the influence of other individual variables as much as possible, as previously done by others authors (e.g. Nichols & Ulatowski, 2007; Mele & Cushman, 2007) and the administration of the version (negative and positive side effects) was balanced to avoid the order effects detected in some studies (Cushman & Mele, 2008; Feltz and Cokely, 2011), as well as the administration of scenarios and ToM and Empathy tasks.

Condition	**Negative side effect**	**Positive side effect**
Text description	The vice-president of a company went to the Chairman of the board and said “We are thinking of starting a new program. It will help us increase profits, and it will also harm the environment”. The chairman of the board answered “I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program”. They started the new program. Sure enough, the environment was harmed.	The vice-president of a company went to the Chairman of the board and said “We are thinking of starting a new program. It will help us increase profits, and it will also help the environment”. The chairman of the board answered “I don’t care at all about helping the environment. I just want to make as much profit as I can. Let’s start the new program”. They started the new program. Sure enough, the environment was helped.
Judgment	*Is the harming of environment intentionally done?*	*Is the helping of environment intentionally done?*
Answer	0 = unintentional - 100 = completely intentional	0 = unintentional - 100 = completely intentional

Table 3. Scenario administered to the participants.

3. Results

In order to assess whether there are associations between demographic variables (age, gender, education) and/or the order of administration with the dependent variable (asymmetry of intentionality judgments), we performed a series of regression analyses: age F(1,85) =.27, β=.05, adjusted R2 =.00, p =.59, gender F(1,85) =.54, β=.08, adjusted R2 = -.00, p= .46, education F(1,85) =.36, β = -.06, adjusted R2=.00, p=.54 and order of administration of scenarios F(1,85) = .35, β=-.06, adjusted R2 =-.00, p =.55, did not predict the asymmetry. For this reason demographic variables were not considered in the subsequent analyses.

Firstly we considered the mean of intentionality judgments provided by participants to the negative side effect and to the positive side effect and the presence of the side effect between the two judgments was confirmed by a t-test analysis t(1,85) = 13.06, p<.001, Cohen's d[7]= 1.8 (see Table 4).

Side effect	***M (SD)*** **Intentionality judgments**
Negative	66.34 (34.32)
Positive	14.53 (18.32)

Table 4 . Means, standard deviations of intentionality judgments to the negative and positive side effects (*p<.001).

In order to assess if the cognitive and affective ToM, controlling for Empathy abilities, predict the ascription of intentionality (see hp 1) a stepwise multiple linear regression was performed considering Empathy, cognitive and affective ToM scores (as independent variables) on the asymmetry of intentionality judgments (as a dependent variable). Multicollinearity

[7] Cohen's d for repeated measure design (Lakens, 2013). Cohen's drm=(Mdiff/sqrt(SD12+SD22-2*r*SD1*SD2))*sqrt(2(1-r)).

diagnostics suggested adequate independence of all predictors (Variance Inflation Factors: 1.00~1.06; Minitab Statistical Software, 2011) (see Table 5).

The model was significant, F(1,84) =22.44, p<.001, adjusted R2=.20. In particular, cognitive ToM predicted the asymmetry of intentionality judgments (β= -.46, p<.001): individuals with a more advanced cognitive ToM ability exhibited a significantly reduction of the judgments' asymmetry. Instead, Empathy, t(1,84) =1.20, p=.23, and affective ToM, t(1,84) =1.15, p=.25, abilities did not predict the asymmetry.

In order to detect which of the two judgments (for the negative and positive side effect) was predicted by the cognitive ToM ability, two separate regression analyses were performed, considering the cognitive ToM score (as independent variable) on the intentionality judgments provided for the negative and for the positive side effect respectively (as dependent variables). The regression analyses indicated that cognitive ToM ability predicted the intentionality judgments only for the negative side effect: individuals with a more advanced cognitive ToM ability provided lower intentionality judgments, F(1,84) =25.55, β= -.48, p<.001, adjusted R^2=.22.

Even though the total Empathy score did not predict the intentionality judgments, we conducted a further series of regression analyses between each subscale of IRI and the asymmetry of judgments and no significant predictions were revealed, F(4,84) =.74; adjusted R^2= -.01; p=.54; PT: t(1,85) = -.76, β= -.08, p=.44; FS: t(1,85) =.08, β=.01, p=.93; EC: t(1,85) =1.49, β=.17, p=.13; PD: t(1,85) =.34, β=.03, p=.73.

			Inter- correlations	
Task	**M(SD)**	**1**	**2**	**3**
1. Short Story task	17.95 (4.15)	-	-	-
2. Reading the mind in the eyes test	23.51 (3.60)	.24*	-	-
3. Interpersonal Reactivity Index	65.83 (10.39)	.06	.11	-

Table 5. Means, standard of deviations and inter-correlations of independent variables. The Short Story task (SST) is from Dodell-Feder et al. (2013); The Reading the Mind in the Eyes Test (RMET) is from Baron-Cohen et al. (2001); The Interpersonal Reactivity Index (IRI) is from Davis (1980). (*p<.05).

Finally, in order to assess the hypothesis 2 concerning whether the cognitive ToM ability was associated with the content of explanation provided for the negative side effect judgment (character’s intention or side effect), participants were divided according to their ToM ability (High:>19 points obtained in the SST task or Low: <18 points obtained to SST task) and their explanations were divided according to the content: concerning character’s intention or the side effect. The Cochran's Q Test for Dependent Samples was significant (Q3=57.57, p<.001). Pairwise comparisons showed that participants (N=46) with a low ToM ability (LToM) provided explanations more focused on the side effect (80.4%, p<.001), while participants with high ToM (HtoM, N=40) provided explanations more focused on the protagonist’s intention (62.5%, p<.05) (see Figure 3).

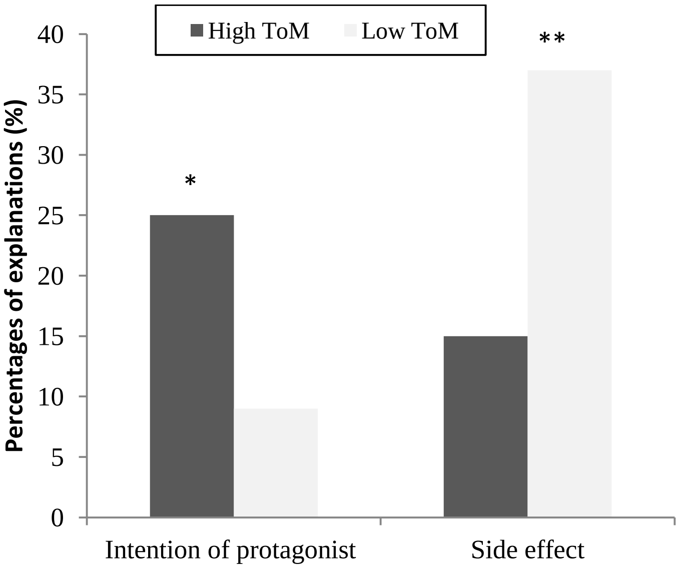

Figure 3. Percentages of explanations provided to the negative side effect judgment, by participants with high or low ToM ability, concerning the intention of protagonist or the side effect. (* p<.05; **p<.001).

4. Discussion

The aim of the present study was to analyze if the individual ability of ToM, controlling for empathy, predicts the ascription of intentionality, using the classic Knobe scenario (2003) which provides an interesting asymmetry of intentionality judgments to side effects in healthy participants.

Firstly, the presence of the *KE* was verified, showing that the negative side effect obtains higher intentionality judgments compared to the positive one.

Since the attribution of intentionality involves the assessment of other people mental states, we analyzed the individual cognitive and affective ToM abilities of the participants who formulated the intentionality judgments. In particular, we were interested in exploring if this individual characteristic can account for a differentiated attentional focus on sundry elements among individuals, which may in turn lead to divergent interpretations and attributions. Indeed in the *KE* individuals are more focused on side effect compared to beliefs and motives of protagonist and we speculate if individual differences in the ability to integrate belief could predict the attribution of intentionality provided. As expected individuals with an higher cognitive ToM ability showed a lower asymmetry between the intentionality judgments provided for the negative and positive side effects, compared to participants with a lower ToM. This result is explained by a decrease in the intentionality attributed to the negative side effect, suggesting that these individuals have paid more attention to the information concerning the intention of protagonist and they embedded it to that about the side effect. This suggestion was confirmed by the analysis on content of participants' explanations, for their negative side effect judgment, which indicated that whose with higher ToM have paid more attention to the protagonist's intention to formulate their decision, whereas individuals with lower ToM ability focused more on the side effect.

In what way? Higher ToM capability allowed to maintain a more robust mental representation of others' beliefs, allowing individuals to focus on the protagonist's intention or motive, and overtaking the emotional/attentional bias caused by the negative side effect.

Likewise, developmental evidence suggests that children with a more advanced ToM ability are more able to make moral judgments, distinguishing accidental transgressions from intentional ones (Fu et al., 2014), and to appropriately attribute responsibility for the behaviour (Hayashi, 2007), when they became able to maintain such a mental representation. In the same manner, adults are more prone to forgive accidental harms in moral evaluations (Patil et al., 2017).

Therefore ToM ability, helping people to integrate the protagonist's intention information, gives rise not only to more elaborate moral judgments but also to a more accurate attribution of intentionality. The Knobe's scenario offers an interesting framework since it shows how people, when the outcome's valence is negative, are mainly "captured" by the side-effect information; from an adaptive point of view the asymmetry attribution shows the key ability of individuals to discriminate between different outcomes according to their valence, but on the other hand it shows that people could be highly “captured” by outcome’s information, ending to neglect that about intention. More generally, this result indicates that our ability of ToM could help us on focusing on intention’s information in contexts in which this is required but other salient information catch our attention or the intentions of an individual are not explicit or ambiguous.

In turn, the capability to take into a greater account the perspective of other individuals has a great impact also on other kind of judgments, such as of responsibility and punishment. In Study 1 we showed that individuals whose actions are perceived to be intentional are also considered more responsible for their action and are assigned with higher length of punishment (Gambetti et al., 2017). Prior studies showed the influence on judgments of blame as well, so that taking the perspective of a criminal defendant, leads to seeing him as less culpable, less guilty and less likely to reoffend, than if you hadn’t, and the same occurs if one takes the victim perspective (Catellani & Milesi, 2001; Galinsky, Ku, & Wang, 2005; Skorinko et al., 2014).

However ToM’s individual differences and intentionality attribution relationship in healthy individuals, is still poorly empirically studied, whereas studies are more focused on patients’ deficits.

We assumed that the cognitive ToM played a principal role on intentionality attribution since, in order to reduce the extent of judgments asymmetry, participants had to analyze the proposed situation focusing on protagonist’s belief, and on its role with respect to the side effect. On the other hand, affective ToM, which represents the ability to understand others’ emotions and Empathy, which entails vicariously sharing others’ affective states, did not show any effect on the attribution of intentionality probably because, in this specific scenario, no information were provided about protagonist’s emotions, therefore participants are not led to analyze the protagonist’s emotions nor to empathize with him. The available information is about protagonist’s intention “to increase company’s profits”. For this reason, we considered to explore in more detail this aspect in subsequent studies, which employ scenarios containing more information about the protagonist’s emotions.

Meanwhile, a recent study provided some tentative answers about this relationship: Slavny & Moore (2017) investigated the association between individual differences in cognitive and affective empathy and the so-called “intentionality bias”. As introduced before, “the intentionality bias” theory (Rossett, 2008) indicates the tendency among people to over-attribute intention when presented with ambiguous actions that could be interpreted as either intentional or unintentional (Moore & Pope, 2014; Peyroux et al., 2014; Rosset, 2008). Using Rosset's (2008) ambiguous sentence paradigm, the authors found that cognitive empathy[8], but not affective empathy, accounted for a significant amount of variance in intentionality bias scores and predicted a higher proportion of intentional over unintentional judgements. According to the authors, this relationship could allow a sort of ‘shared intentionality’ among humans, whereby people participate in collaborative activities involving shared goals and intentions (Tomasello et al., 2005).

In conclusion, the present results provide evidence that the assessment of individual differences can be an essential tool in understanding intentionality attribution, suggesting the necessity to take more into account the significant individual differences in the ability to evaluate social situations and to explain others’ behaviour. Study 2 showed that people can focus on different information (intention or side effect), while providing judgments, according to their individual abilities.

[8] Measured with the QCAE (Reniers et al., 2011): a self-report questionnaire that examines the respondent's ability to understand the emotional states of others (cognitive empathy) and their ability to vicariously experience what others are feeling (affective empathy).

Experiment 3: The influence of Autistic traits of personality on attribution of intentionality in typically-developing individuals

1. Introduction

Study 2 underlined a significant influence, on intentionality attribution, of individual abilities in Theory of Mind. Specifically, a greater cognitive Theory of Mind (ToM) ability allows participants to focus more on intention information of Knobe scenario, allowing them to decrease the intentionality attributed to the negative side effect, resulting in a reduced asymmetry. On the other hand, participants with a lower theory of mind ability focused more on the side effect information and therefore showed higher intentionality judgments and a greater asymmetry.

This result suggested that the individual ability of Theory of Mind addresses our attention on different element, in the social context and this has, in turn, a great effect on intentionality attribution.

Since Theory of Mind ability was found closely associated with intentionality attribution, we decided to further analyze this relationship, exploring intentionality attribution in a sample of typically-developing individuals who present autistic traits of personality.

Indeed ToM ability is notoriously diminished in individuals with autism spectrum disorders (ASD- APA, 2014), however, crucially, recent findings outlined that performance of typically-developing individuals with high levels of autistic traits are associated with atypically ToM abilities as well (Gökçen et al., 2014, 2016; Lockwood et al., 2013). In particular the authors found atypically perspective-taking abilities on the *animated triangles* task (Abell, Happé & Frith, 2000), in decoding mental states from the eye region at *Reading the Mind in the Eyes* task (Baron-Cohen et al., 2001a) and greater difficulties in identifying mental states from dynamic video-based stimuli (*Movie for the Assessment of Social Cognition*- MASC; Dziobek et al., 2006). Indeed, there is growing evidence that the expression of sub-threshold ASD traits may extend into the general population (Baron-Cohen et al., 2001b; Constantino, 2011; Hoekstra, Bartels, Cath, & Boomsma, 2008; Jones, Scullin, & Meissner, 2011). In particular previous studies (e.g. Lockwood et al. 2013) have found an association between autistic traits and reduced cognitive perspective-taking (Theory of Mind), but not affective resonance (empathy) which has shown instead to be intact, converging with autism literature (Jones et al. 2010; Schwenck et al. 2012).

Therefore, according to this recent literature, the aim of Study 3 has been to explore if the presence of autistic traits of personality, in typically developing individuals, could influence the attribution of intentionality to a side effect. Moreover we wanted to investigate if the relationship between autistic traits of personality and intentionality attribution is mediated by theory of mind abilities (in order to replicate results obtained in Study 2 as well).

Indeed some studies have showed intentionality attribution difficulties in Aspergers (HFA) and ASD individuals but, to our knowledge, no investigations were performed in typically developing individuals with autistic traits of personality and considering side effects. In particular, Zalla et al. (2009) analyzed the ability of intentionality attribution in individuals with HFA syndrome, finding an over-attribution of intention in the *Faux Pas* task, which measures the ability of individuals to recognize the unintentional nature of mistakes done by the protagonists of the stories (Stone et al., 1998). In the study, HFA individuals, as the control group, detected socially inappropriate behavior but they incorrectly attributed to the speaker the intention to hurt the listener and sometimes even judged the former to be blameworthy. The same behavioral pattern was found by Buon et al. (2013) who underlined that individuals with HFA showed difficulties in judging the agent’s intention: they tended to overrate the agent’s intention to harm in the accidental conditions and consequently regarded him as being more responsible and they punished him more severely than control participants.

1.2 The present study

The aim of Study 3 was to investigate the attribution of intentionality to a side effect in a sample of typically developing individuals who present high levels of autistic traits of personality.

The first hypothesis was to confirm the association between autistic traits of personality and lower ToM abilities in a group of individuals who exhibit this trait of personality, compared to a control group, according to the recent literature (Gökçen et al., 2014, 2016).

Consequently, the second hypothesis was that typically developing individuals with autistic traits of personality would show also atypical intentionality attribution to a side effect; indeed according to Study 2 lower theory of mind abilities lead to increased intentionality judgments for the negative side effect in the Knobe scenario; consequently higher intentionality attribution is expected for this kind of scenarios in individuals with autistics traits, who exhibit reduced theory of mind abilities (Gökçen et al., 2014), compared to a control group without this trait. As regards ToM ability, Study 2 showed a greater involvement of the cognitive subcomponent of ToM in the intentionality attribution process, with respect to the affective one. However, a tentative explanation provided about this result was focused on the specific task used (the classic Knobe scenario) which could have required only cognitive ToM due to the reduced presence of emotional cues. In order to clarify this result, the present study employed more scenarios, like the classic one, but characterized by a greater emotional content, and so expected to involve both ToM subcomponents. Moreover, in the literature on ASD, some studies suggested that 'cognitive' and 'affective' ToM abilities are differentially impaired, but the results have been inconsistent, stating that only one component was impaired or instead both of them (e.g. Shamay-Tsoory et al., 2002; Dziobek et al., 2008; Mazza et al., 2014), while any result has been yet collected in typically developing individuals with autistic traits of personality.

Finally, in order to confirm the relationship between autistic traits of personality and intentionality attribution to a side effect, the mediation role of ToM abilities was explored: the third hypothesis suggested that theory of mind abilities, associated with autistic traits of personality, mediates the relationship between the latter and intentionality attribution.

2. Method

2.1. Participants

To determine the sample size, a series of power calculations were performed, using GPower 3.1. (Faul et al., 2007). First the sample size necessary to perform a mediation analysis, considering autistic traits of personality and theory of mind on intentionality attribution, was carried out (effect size f^2= .15; alpha = .05; power = .90). This gave a suggested sample size of 99 participants.

Secondly, we planned to compare the group with autistic traits to a control group. Therefore an additional power calculation for a repeated measure Anova (effect size d= .25; alpha = .05; power = .90) suggested a sample size of 46 per group. According to these power calculations, the total sample was set around 100 participants. Hundred and four volunteers were recruited at the University campus through notices on social networks and on bulletin boards. The exclusion criteria included a history of neurological or psychiatric disorders. Four participants were excluded because they met the exclusion criteria.

Due to four excluded participants the final sample size was 100 participants (47 males, students and young workers), aged between 20 and 40 years (M= 24.00, SD=5.21), education (20% master degree, 24% graduate degree, 56% secondary/high school). All participants were native Italian speakers and had normal or corrected-to-normal vision.

All participants signed a written consent form before the study began. The study was approved by the local Ethics Committee of the University of Bologna.

2.2. Materials and Procedure

2.2.1 Scenarios

Participants were presented with fourteen scenarios, each composed of a negative and a positive consequence variant-modelled after the original vignettes, taken from Ngo et al. study (2015), for a total of 28 variants[9]. The English version was translated before into Italian by an Italian native speaker, proficient in English. Then, the translated dilemmas were presented to an English native speaker proficient in Italian for the back-translation to English (Brislin, 1970). Two examples of the scenarios are shown in Table 6. The others scenarios are reported in Appendix A. Each variant presented, as the classic Knobe scenario, an individual who performs an action, which will be associated with a negative or a positive side effect respectively. Participants are requested to determine the level of intentionality of the side effect using a scale from 0 (unintentional) to 100 (completely intentional).

For the statistical analyses, the mean of intentionality judgments provided for the fourteen negative variants was computed to obtain a global score; the same was carried out for the positive side effects.

As in Study 2, we followed a within-subjects design, whereby participants had to face each scenario in the positive and the negative version (e.g. Nichols & Ulatowski, 2007; Mele & Cushman, 2007) and administration of the version (negative and positive side effects) was balanced (as in Cushman & Mele, 2008; Feltz and Cokely, 2011).

[9] The original set (of 80 scenarios) was reduced to not induce too much fatigue for the participants. The reliability of scenarios selected was previously verified through a pilot study.

Condition	**Negative side effect**	**Positive side effect**
Example 1 *Text* *description*	Roger enacted a financial scheme to buy a house. Roger did not care at all about the effect the scheme would have on old retirees. Roger knew his plan would bankrupt old retirees.	Renee enacted a financial scheme to buy a car. Renee did not care at all about the effect the scheme would have on old retirees. Renee knew her plan would help old retirees.
Judgment	*Did Roger intentionally bankrupt old retirees?*	*Did Renee intentionally help old retirees?*
Example 2 *Text* *description*	The airplane bomber bombed a factory to reduce enemy's steel production. He did not care at all about the effect the bombing would have on innocent civilians. He knew his bombing would kill innocent civilians.	The bomber pilot bombed a facility to reduce the enemy's iron production. She did not care at all about the effect the bombing would have on the townsfolk. She knew her bombing would liberate the townsfolk.
Judgment	*Did the airplane bomber intentionally kill innocent civilians?*	*Did the bomber pilot intentionally liberate the townsfolk?*
Answer	0 = unintentional - 100 = completely intentional	0 = unintentional - 100 = completely intentional

Table 6. Examples of scenarios

2.2.2. Questionnaires

In order to measure the level of ToM abilities of participants (see Hp1), as in Study 2, *Reading the Mind in the Eyes Test* (RMET, Baron-Cohen et al., 2001) and the *Short Story Task* (SST, Dodell-Feder, Lincoln, Coulson & Hooker, 2013) were employed, due to their properties in detecting individual differences in healthy individuals (e.g. Dodell-Feder et al., 2013). For more detailed information see Study 2.

Autism and Asperger Diagnostic Scale Revised. In order to measure the presence of autistic traits of personality, participants filled in the Ritvo Autism and Asperger Diagnostic Scale

Revised (RAADS-R, Ritvo et al. 2011). The scale is composed of 80 items, which comprises four symptom areas: language, social relatedness, sensory-motor and circumscribed interests. Individuals have to indicate, among four options "now and when I was younger", "only now", "only when I was young", "never", how much each item applies to them. The four options allocated 3, 2, 1 and 0 points respectively and have to be summed to obtain a total score, which ranges from 0 to 240 points. The cut-off suggested by the authors is above 65 points (Ritvo et al. 2011).

The means, standard deviations, and bivariate correlation coefficients for all variables can be seen in Table 7.

Task	***M***	***SD***	**1**	**2**	**3**
Short Story task	15.99	2.80	-	-	-
Reading the mind in the eyes test	23.62	3.43	.18*	-	-
Autism and Asperger Diagnostic Scale Revised	61.98	25.78	-.25**	-.26**	-

Table 7. Means, standard deviations and bivariate correlation coefficients for all variables. The Short Story task (SST) is from Dodell-Feder et al. (2013); The Reading the Mind in the Eyes Test (RMET) is from Baron-Cohen et al. (2001); Autism and Asperger Diagnostic Scale Revised (RAADS) is from Ritvo et al. (2011). (* p<.05; ** p<.001).

2.2.3. Procedure

All participants were individually tested by a single researcher in a quiet room for about an hour. Instructions were provided at the beginning of each test. Firstly they are required to complete a demographic questionnaire about their age, gender and education level, then progressed through the self-report measures that assessed variables of interest and, at last, answered Knobe-like scenarios.

3. Results

3.1 Correlations between autistic traits of personality and theory of mind abilities

First a correlation analysis was performed to verify Hp1: the association between high autistic traits of personality and atypically ToM abilities in typically developing individuals. The results showed a negative correlation between RAADS-R scale e RMET task score [RAADS-R – RMET: r (100) = -.26, p<.05] and between RAADS-R scale and SST task score [RAADS-R –SST: r (100) = -.25, p<.05]. With increase of autistic traits of personality, ToM abilities of participants decrease.

3.2 Intentionality judgments to side effects and autistic traits of personality

The mean of the judgments provided to the fourteen scenarios, calculated to obtain a global score, for the negative side effects was compared to the mean obtained for the positive side effects. The presence of the *Knobe Effect* (Knobe, 2003) was verified, in the overall group, by a paired t-test analysis ($t_{1,99}$ = 12,02; p<.001; IC: 23.23-32.41; Cohen's d[10] = 1,02). Descriptives are in Table 8. Participants judged the negative side effects as being significantly more intentional with respect to the positive ones.

[10] Cohen's d for repeated measure design (Lakens, 2013).

Side effect	**M (SD) of Intentionality judgments**
Negative	68.37 (20.31)
Positive	40.55(18.71)

Table 8. Mean and standard deviation of intentionality judgments provided to negative and positive side effects of scenarios (**p<.001).

Subsequently the sample was divided, according to the RADDS-R scale cut-off, into two groups[11]: participants who present autistic traits of personality (N=48; age: 23.20(5.72); 24 males) and the control group (N=52; age: 24.15(4.23); 23 males).

In order to assess the association among demographic variables (i.e. age, gender, education) and the order of administration of scenarios (1=negative first; 2= positive first) on the dependent variables, a series of correlation analyses were performed for each judgment (i.e., intentionality judgments for the negative and positive side effect): age (r=.00, *p*=.94), gender (r=.04, *p*=.66), education (r=-.07, *p*=.48) and order (r=.-11, *p*=.28) weren't correlated to the negative judgment, nor to the positive one: age (r=.14, *p*=.16), gender (r=.03, *p*=.71), education(r=-.08, *p*=.39) and order (r=.-11, *p*=.27), so they are not considered in the subsequent analyses.

[11] The inclusion of participants was also verified submitting them the Autism Spectrum Quotient (AQ-10 Baron-Cohen et al., 2001b, see Appendix B) in a different session.

In order to verify Hp2, which predicts higher intentionality judgments for the negative side effect exhibited by individuals with autistic traits of personality, a 2 valence (negative and positive side effects) X 2 groups (individuals with and without autistic traits of personality) repeated measure Anova was performed, considering the mean of the intentionality judgments provided for each side effect.

The analysis revealed a main effect of valence [F(2,98)= 143.22, p<.001, η^2=.59]: intentionality judgments provided for the negative side effects were significantly higher than those provided for the positive ones and a main effect of group [F(2,98)= 9.26, p<.05, η^2=.08]: participants with autistic trait of personality provided higher intentionality judgments with respect to the control group (Figure 4).

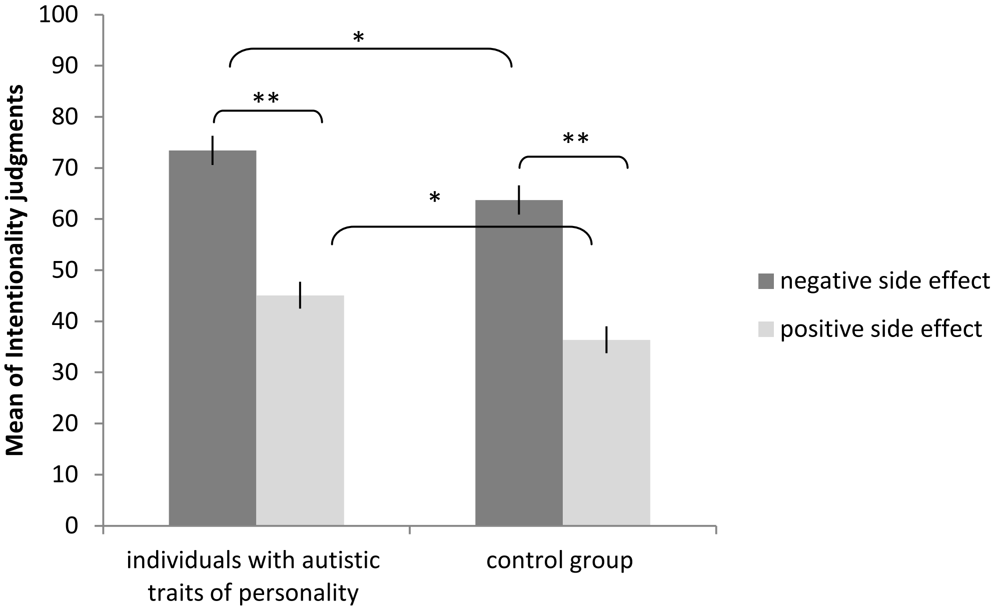

Figure 4. Mean and standard error of intentionality judgments provided by typically developing individuals with autistic traits of personality and the control group. Significant differences are indicated as follows: *p<.05; **p<.001

3.3 The mediation role of theory of mind in the relationship between autistic traits of personality and intentionality judgments

In order to verify Hp3, that ToM abilities mediates the relationship between autistic traits and intentionality judgments, we tested the mediation effect of cognitive and affective theory of mind abilities by the approach laid out by Baron and Kenny (1986) that requires four steps through regression equations.

The first step tests for a significant relationship between the independent (i.e. *autistic traits of personality*-RADDS) and the dependent variable (i.e. *judgments of intentionality for the negative and positive side effects*). The second step tests for a significant relationship between the independent variable (i.e. *autistic traits of personality*-RADDS) and the mediator (i.e. *theory of mind ability*: cognitive-SST=1; affective-RMET=2). The third looks at the relationship between the mediator and the outcome variables (i.e. *judgments of intentionality for the negative and positive side effects*). If these equations are significant, a fourth regression equation is computed in which the independent and the mediator variables are included as predictors of the outcome variables. Mediation would be either full or partial. Full mediation can be inferred if the regression coefficient for the predictor is not significant in the fourth regression equation, while partial mediation can be inferred if the value drops, but remains significant. Significance of mediation is evaluated using the Sobel test (Sobel, 1982).

In the first test for mediation, the regression analysis revealed a relationship between autistic traits of personality and intentionality judgments for the negative side effects (β=.29, p < .05) accounting for 8% of the variance, and between autistic traits of personality and intentionality judgments for the positive side effects (β=.28, p < .05), accounting for 8% of the variance.

In the second test for mediation, the regression analysis revealed a relationship between autistic traits of personality and cognitive theory of mind (β = -.25, p < .05), [SST: F(1, 99) = 7,02, p < .05], accounting for 6% of the variance, and between autistic traits of personality and affective theory of mind (β = -.26, $p < .05$), [RMET: F(1, 99) = 7.18, $p < .05$], accounting for 6% of the variance.

In terms of the dependent variables, we performed two regression analyses considering first intentionality judgment for the *negative* side effect and then intentionality judgments for the *positive* one. So in the third test for mediation, the regression analyses showed a predictive effect of cognitive theory of mind on intentionality judgment for the *negative* side effect (β = -.34, $p < .001$), [F(1,99) = 12.83, $p < .001$], accounting for 11% of the variance; and for the positive side effect (β = -.22, $p < .05$), [F(1,99) = 5.06, $p < .05$], accounting for 4% of the variance. Instead, affective theory of mind ability doesn't predict intentionality judgments for the *negative* (p=.15), nor for the *positive* side effects (p=.78).

In the fourth test for mediation, we examined whether the relationship between autistic traits of personality and intentionality judgments for the side effects was attenuated when the mediator, that is cognitive theory of mind, was included in the regression model. The analysis showed that intentionality judgments for the *negative* side effect were associated with both autistic trait of personality (β = .22, $p < .05$) and cognitive theory of mind ability (β = -.28, $p < .05$) [F(2, 98) = 9.33, $p < .001$]. Autistic traits of personality accounted for 8% of the variance in the score of intentionality and cognitive theory of mind ability explained another 8% .

Then the analysis showed that intentionality judgments for the positive side effect was only associated with autistic traits of personality (β = .24, $p < .05$) and no with cognitive theory of mind ability (p=.11) [F(2, 98)= 5.59, $p < .05$]. Autistic traits of personality accounted for 10% of the variance in the score of intentionality.

Calculation using the Sobel test indicated that cognitive theory of mind ability (Z = 2.03, $p<.05$) partially mediated the relationship between autistic traits of personality and intentionality judgments for the *negative* side effect (Figure 5).

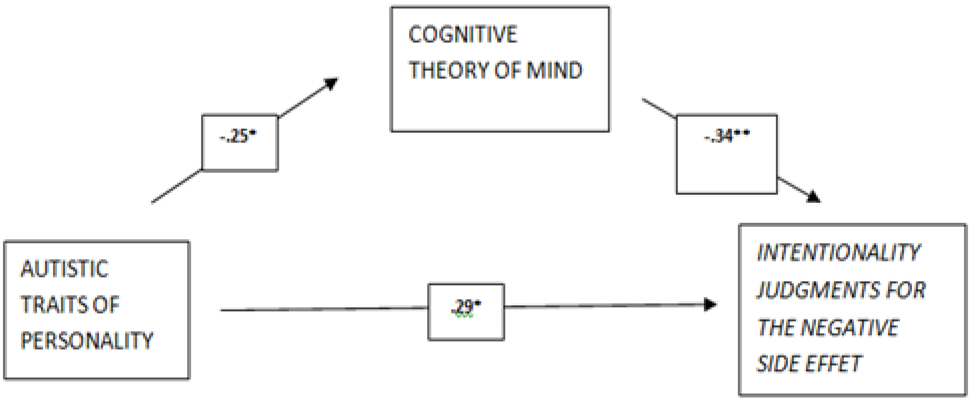
COGNITIVE THEORY OF MIND
-.25*
-.34**
AUTISTIC TRAITS OF PERSONALITY
.29*
INTENTIONALITY JUDGMENTS FOR THE NEGATIVE SIDE EFFET

Figure 5. Mediation model showing the partial mediation effect of cognitive theory of mind between autistic traits of personality and intentionality judgments for the negative side effects. (*p<.05; ** p<.001).

4. Discussion

Recent studies have shown that typically developing individuals with autistic traits of personality can exhibit ToM difficulties, milder but similar, to ASD (Gökçen et al., 2014, 2016). Due to the close relationship between ToM abilities and the attribution of intentionality found in Study 2, and to the intentionality attribution difficulties exhibited by ASD individuals (e.g. Zalla et al. 2009), the aims of Study 3 were: to verify if autistic traits of personality in typically developing individuals are associated with atypical ToM abilities (Hp1), to analyze if the attribution of intentionality to a side effect was also affected in individuals with autistic traits of personality (Hp2) and finally to verify if the relationship between autistic traits of personality and intentionality attribution is mediated by theory of mind abilities (Hp3).

Firstly the results of the present study confirmed recent findings (Gökçen et al., 2016) about reduced ToM abilities associated with autistic traits of personality in typically developing individuals, similar (although milder) to that of ADS individuals. However in the literature there are no studies which investigated individually both components of ToM (Shamay-Tsoory et al., 2005). Therefore, we explored the performance on each one, finding reduced both cognitive and affective component.

Then we explored the relationship between autistic traits and intentionality attribution. Study 2 showed that individuals with higher ToM ability pay more attention to the protagonist's intention and therefore they reduced the intentionality attributed to the negative side effect while, on the other hand, individuals with a lower ability pay more attention to the side effect information to formulate their judgments, thus increasing the intentionality attributed to the latter. The present study, with a greater variability of scenarios, confirmed the relevance of ToM ability on intentionality attribution, and to partly explain the *KE*: indeed, higher presence of autistic traits of personality, and consequently lower ToM abilities, increases the attribution of intentionality; moreover the attribution is higher not only for the negative side effect, as expected, but for both side effects; we can therefore hypothesize that the attention of typically developing individuals with autistics traits of personality is more drawn to the side effects, mostly ignoring the protagonist's intentions.

Finally, to better understand the relationship between autistic traits of personality and intentionality attribution (Hp3) we performed a mediation analysis, which confirmed the partial mediation effect of theory of mind, in particular the cognitive one, on intentionality judgments provided for the negative side effect. Instead, the relationship between autistic traits of personality and intentionality judgments for the positive side effect was not partially mediated by cognitive theory of mind and there was a direct relationship with autistic traits of personality. As a matter of fact, Study 2 showed the same result: cognitive theory of mind ability only predicted the intentionality judgments for the negative side effect. However, in this instance, judgments about the positive effect formulated by individuals with autistic traits, were also higher compared to the control group and to what occurs in the classic *KE*. We attempt to speculate about this result: it is possible that the increase of intentionality judgments for the positive side effect was driven not only by the lower ToM abilities of individuals with autistic traits, but also from other features typically belonging to the autistic trait pattern, such as repetitive behavior and poorer cognitive flexibility (e.g. South, Ozonoff & Mcmahon, 2007; Gökçen et al., 2014), which may have pushed them toward a "sort of generalized approach to the judgments", regardless of the specific stimulus. In fact, in individuals with typical development, information about the agent's innocent intention is loaded on executive resources, to inhibit the strong tendency to blame an agent for the harmful outcome of his action, and mitigate the blame by reason of the agent's innocent intention (Young et al. 2007). Moreover, theory of mind ability, which is noticeably reduced in this sample, is necessary to correctly focus and attribute intentions, compared to outcome information (in this case the side effect) (Young et al. 2007; Patil et al., 2017), so the attribution of intentionality to the positive side effects may have suffered this shortcoming too.

To summarize, the presence of autistic traits of personality increases intentionality judgments provided to the side effects (negative and positive) and cognitive theory of mind ability partially mediates this relationship (with increase of cognitive ToM, the intentionality attribution to the side effect decreases) for negative side effects. Affective theory of mind failed again (as in Study 2) to predict intentionality judgments, beside the greater variability of scenarios used.

Despite the higher intentionality judgments provided for both side effects by the group with autistic traits of personality, the latter exhibits the *KE* in any case, as the control group. They attributed higher intentionality to the negative side effects compared to the positive ones. The prominence of this effect could probably be explained by the negative consequences of the first ones, which provoke an emotional reaction in the reader, as confirmed by a recent fMRI study (cited in the *Blame and Affective hypothesis* paragraph; Ngo et al., 2015), which found higher levels of amygdala activation associated with higher intentionality judgments for the negative side effect. The partial involvement of amygdala in the increasing of intentionality attribution is consistent with studies which posited the presence of an amygdala hyper-activation, resulting from a defective top-down modulation by prefrontal areas, in ASD (e.g. Zalla & Sperduti, 2013). On the other hand, we could expect a reversed effect in a group of individuals who present a reduced amygdala activation, such as individuals with alexithymic traits of personality (Reker et al., 2010), or even more patients who present a selective amygdala lesion and so an overall lack of activation. This hypothesis has been explored in Study 4 in a group of individuals who exhibited alexithymic traits of personality.

The results of Study 3 added a step to the investigation about the difficulties exhibited from individuals with autistic traits of personality, showing that their ToM difficulties are also reflected in the intentionality judgments provided to a side effect. Although in the subclinical context, outlining performance differences between individuals with autistic traits of personality and without them, represents a great concern. Previous studies already detected poor performances on attributing mental states to movie characters in a real-life social context, emotional intelligence and cognitive flexibility in this population (Gökçen et al., 2014, 2016). Moreover, prior studies showed a significant role of subclinical traits of personality on intentionality attribution, such as in individuals affected from schizotypy (Moore & A. Pope, 2014) who exhibit a stronger intentionality bias, according to Rossett's paradigm (2008).

Instead, as regards autism spectrum, to our knowledge, prior studies explored intentionality attribution only in HFA and ASD individuals, showing that when judging accidental harms, participants who successfully passed a standard false belief task, exhibited an under-reliance on information about a person's innocent intention, together with an over-reliance on the action's negative outcome (Moran et al., 2011), as occurred for the side effects in our study. According to the authors, these findings reveal impairments in integrating conflicting information about mental state (e.g. neutral intentions) and (aversive) action outcome in moral judgment, hypothesizing that the difficulty in conciliating conflicting information about an agent's intention and the action outcome is due to the lack of a robust and flexible ToM, necessary to override the overpowering response driven by emotionally salient information. The presence of an overbearing response driven by emotionally salient information could be sound atypical in relation to autistic traits of personality but it is not disagree according latest models on ASD (e.g. Bird & Cook, 2013; Rogers et al., 2007; Smith, 2009; Lockwood et al., 2013), which clearly distinguished between cognitive (meaning *the whole theory of mind ability*- intended as the ability to understand others' thought and emotions) and affective deficits (meaning *empathy ability* - intended as the ability to share others' emotional states - Singer, Critchley, & Preuschoff, 2009) in this population, proposing a greater responsiveness exhibited towards others' emotional states. Rogers et al. (2007) found no impairments in the *Empathic Concern* scale and higher scores on the *Personal Distress* scale of Interpersonal Reactivity Index for Empathy (IRI - Davis, 1980) than controls, indicating a greater tendency to have self-oriented feelings of anxiety and discomfort in response to difficult interpersonal contexts. They suggested, also in keeping with anecdotal reports from parents and clinicians revealing that autistic individuals can be very caring, that individuals with ADS appear to have as much care and concern for other people as unaffected individuals do. On the other hand, low scores on the *perspective taking* scale of the IRI (which assess the tendency to spontaneously adopt the psychological point of view of others in everyday life) could explain why ASD do not react to situations as expected and consequently why they seem to be cold or uncaring (Rogers et al., 2007). Smith (2009), Jones et al. (2010) and Schwenck et al. (2012) reported difficulties in mentalization but an intact capacity to resonate with another person's emotions; which is consistent with a greater responsiveness to others' emotional states exhibit by children with ASD as well (Capps et al. 1993). Dziobek et al. (2008) also found no group differences in the affective domain of their *Multifaceted Empathy Test*, while clear deficits in the ability of ASD to infer another person's mental state. Moreover ASD judge accidental harms more harshly, due to their inability to form a robust representation of agent's benign intentions due to ToM deficits (Leekam, 2016), and they exhibited increased tendency to reject the utilitarian option, in emotionally salient dilemmas that required direct physical harm to a victim (e.g. pushing someone to their death), finding all conditions (non moral, personal and impersonal moral dilemmas) to be more arousing than controls (Patil et al., 2016). Indeed, in the same study, after accounting for co-occurring alexithymia, autism is no longer associated with deficits on dispositional empathy (see also Aaron, Benson & Park, 2015). Finally autistic children also exhibit typically physiological responses to others' pain (Fan et al., 2014; Hadjikhani et al., 2014) and distress (Blair, 1999). Consistently, we found reduced theory of mind abilities in individuals with autistic traits of personality, whereas we did not explore empathy abilities, since they are not related to the present study's purposes (see Study 2 which did not reveal any relationship with intentionality attribution), and they are supposed to be intact (or increased) accordingly to ASD literature. However further studies could explore this feature in more detail, given recent studies (see Slavny & Moore, 2017).

Coming back to our results, it is important to note that a previous study by Zalla et al. (2011) investigated the intentionality and morality judgments in HFA, but found no significant differences in the two groups concerning intentionality judgments. This may sound strange since the selected participants are clinician with respect to our sample, but we suppose these different results could be driven at least by two relevant methodological differences: first of all Zalla et al. (2011) employed only two vignettes while we employed a larger set of vignettes (28 vignettes from Ngo et al., 2015) which involve an high stressor on the victim (which could have increased the emotional response of individuals with autistic traits, according to their great responsiveness to others' stress); second the rating scales employed: Zalla et al. (2011) employed a binary answer ("yes or no") versus our hundred rating scale, which probably has captured more individual variability. Anyway, the average of intentionality judgments provided by HFA participants, in Zalla et al. study (2011), was very higher mostly for the negative side effect (negative side effect: 100% and positive side effect: 36,8%), in accordance with our results. Moreover, participants in the study are also requested to provide moral judgments, and HFA also provided higher judgments of praise for the agent with respect to control group (83.3%). When asked to justify their judgments, 75% of them pointed to the goodness of the action outcome, ignoring the lack of protagonist's intention, while intentions were consistently cited during justification provided by control group. Therefore, while the majority of control group exhibited relevant ToM competencies in understanding the agent's intention, moral judgments about agent's action in individuals with HFA appeared to be primarily affected by the goodness/badness of the outcome (Zalla et al., 2011). Indeed, previous evidence has shown that individuals with HFA are often found to be severely impaired in spontaneous intuitive on-line computation of others' mental states (Senju et al. 2009) and, as mentioned before, they mistake in attributing the intention to the agent in the *Faux Pas* task so that harmful side-effects are regarded as intentional by them (Zalla et al., 2009).

Due to increasing evidences about the relationship between autism and alexithymia (e.g. Aaron et al., 2015; Bird et al., 2013, 2010) a limit of our study could be to have not controlled for co-occurred alexithymia influence. However, as introduced before, we were interested in the ToM deficits exhibited by typically developing individuals with autistic traits of personality, which could be linked to intentionality attributions, and not to empathy ones (which characterize these individuals) therefore we did not take into account this feature in this study. Anyway, this personality trait was explored in Study 4, to investigate the role of individual differences in decoding emotions on intentionality attribution.

Another interesting related feature, to further investigate, and to consider in future mediation analysis in association with ToM abilities, could be the influence of executive functions (for instance "higher working memory ability in retain both scenarios in mind") on intentionality attribution and in particular on the *KE,* due to the connection with autistic traits of personality: indeed Gökçen et al. (2014) found that typically developing individuals with high autistic traits have poorer shifting efficiency and show more perseverative errors, as well as ASD individuals (Ozonoff, 1997; Pennington and Ozonoff,1996; Hill and Bird, 2006).

Importantly, the present results outlined that cognitive ToM must represent a key process to consider when designing intervention programs targeting adaptive social functioning in typically developing populations with elevated levels of autism traits, who showed to increase the attribution of intentionality towards both negative and positive side effects. This is especially relevant given the estimated huge worldwide incidence of autism spectrum disorder ("*in 2010 there were an estimated 52 million cases of ASDs, equating to a prevalence of 7.6 per 1000 or one in 132 persons*" Baxter et al., 2015).

Interventions within the interpersonal sphere typically focus on broader, more goal-oriented aspects of social interactions (improving general conversational skills, interpersonal relationships) and their application to real-world settings. However, given the significant overlap between social and non-social domains of cognition, it will be important to address first to more 'basic' cognitive processes, for instance focusing on theory of mind abilities with specific training, such as the metacognitive education (Gvozdic et al., 2016) which showed reliable effects, to support subsequently more broader approaches, targeting higher-order social competencies.

Experiment 4: Alexithymic traits of personality and intentionality attribution: the role of individual differences in processing emotions

1. Introduction

1.1 *The role of emotions in the Knobe Effect*

The last study of the present dissertation explored the influence of individual ability in processing emotions on intentionality attribution. Prior studies (e.g. Malle and Nelson, 2003) hypothesized a role of emotions in increasing intentionality attribution in the *KE*, however, the results collected were both in favour of and against it, and so they only partially supported this hypothesis.

The hypothesis of the role of emotions in influencing attribution of intentionality was posited for the first time by Alicke (Alicke, 2008; Alicke and Rose, 2012) in his *motivational bias account.* In his model, the author posited that the tendency to blame the agent in cases of negative side-effects can bias the subsequent attributions of intentionality. In turn, attributions of blame depend from the perceptions of the person who judges, about the degree of "control" exerted by the agent on the situation. According to the model, this perception of "control" is determined by three factor: the mental states of the agent about the possible consequences of his behaviour, the behaviour undertaken, and the consequences of that behaviour. For instance, the awareness of the danger and prohibition of drinking before driving (mental states); the decision to drive in any case (the behaviour) and the occurrence of an accident (the consequence).The stronger the relationship between these three elements, the more people perceive a "control" over the situation on the part of the agent, and for this reason they attribute more blame to him. On the other hand, if the relationship between these elements is somehow constrained, the evaluation of the agent's control over the situation is reduced and, in turn, the blame assigned to him; for example, if the agent decides not to drive, but one friend asks him for a ride because he drank more, which the agent agrees to do, and as a consequence they have an accident. Importantly, Alicke (2008) postulated a possible psychological mechanism that trigger this tendency to blame the agent: *spontaneous affective reactions* to the nature (negative or positive) of three elements themselves and to other ‘external-factors’ such the agent’s appearance, reputation, social status, etc. In the case of negative reactions, people who judge could review the link between the three elements (mental states, behaviour and consequences) in a biased manner, by exaggerating their strength or de-emphasizing exculpatory evidence. This cognitive operation is called “blame validation mode”. Therefore according to Alicke, in the “Harm situation of the Knobe scenario”, negative evaluations of the chairman’s attitude and about the outcome of his action, make people perceive more “control” on the part of the Chairman over the situation (by perceiving a stronger relationship between his mental states and the consequences that occurred); for this reason, they attribute to him a greater blame and, in turn, a greater intentionality to his actions (Alicke & Rose, 2012). Another common example describes a driver who had an accident while speeding, alternatively to hide an anniversary present or because of a vial of cocaine, and he encountered a number of environmental obstacles (slippery road, poor visibility). Individuals are more prone to attribute the accident to the driver, rather than the environmental conditions, when the driver was hiding the cocaine, as opposed to when he was hiding an anniversary gift, highlighting that “spontaneous evaluations of the agent’s motives lead participants to exaggerate his control over the situation” (Alicke, 2000).

The same “blame validation operation” could occur, according to Nadelhoffer (2004b, 2006), when a juror has to evaluate a defendant whose actions are immoral. The immoral nature of his actions might trigger negative reactions in the evaluator which, in turn, prejudices his assessment of the defendant’s “control” over the situation. The juror could perceive a greater “control” on the part of the defendant, looking selectively for evidence that supports blame attribution or overlooking factors that might mitigate blame. Indeed, in one of his experiments (Nadelhoffer, 2006) the case of a policemen knocked down by a car driven by a thief, as a side effect of his escape attempt after a robbery, was compared to the same outcome produced by another man, who was going home to his family. The former action is considered more intentional by participants compared to the latter, despite the situational features being the same.

So, the moral blameworthiness of an agent biases the judgments of intentionality regarding his action or its side effects. According to Malle and Nelson too (2003), this is what occurred in the Knobe scenario: the negative affective reaction to the Chairman in the harm scenario induced people to consider the side effect produced as intentional. Indeed, according to appraisal theories of emotion (e.g. Weiner, 2001), every emotion comes with a set of implicit beliefs about oneself, the other person, or the situation; and anger in particular is often associated with the appraisal of the other person's intentional agency. In a case of a fight between two individuals, the anger elicited from the conflict might induce greater beliefs that the other person is doing actions with intent, as in the case of the Chairman, where anger elicited from negative consequences could lead to the perception that his actions were intentional (Malle and Nelson, 2003). Indeed, previous studies showed that inducing anger in jurors during a capital punishment trial simulation led to an underestimation of mitigating circumstances and increased the probability of assigning a death sentence (Georges et al., 2013; Nuñez et al., 2015); increased attributions of causal control by the agent (Ask and Pina, 2011); and increase higher blame attributions (Lerner et al., 1998; Goldberg et al., 1999).

However, a recent experiment (Diaz et al., 2017) tested the hypothesis of the key involvement of anger in determining the KE, but finding no results. The authors measured the participants' tendency to feel anger, as well as inducing anger in them before answering questions related to the scenario, but in both cases they failed to prove the involvement of anger in driving the *KE*. Therefore, they claimed that the underlying factor is non-emotional, but instead is just the tendency to blame the Chairman for the consequences of his action. This conclusion is due to the fact that judgments concerning the responsibility of the Chairman were consistently positively correlated with intentionality attributions (the higher responsibility assigned to him, the higher judgment of intentionality provided).

Another result opposed to the “emotional hypothesis” of the KE was also obtained by Young et al. (2006), who found that patients with a lesion in the ventromedial prefrontal cortex (VMPFC) exhibited the asymmetry of intentionality judgments as well, despite their typical difficulties in emotional processing. The same result was obtained by Cardinale et al.(2014) with psychopathic individuals. However, it is important to note that the emotional deficits of VMPFC patients and psychopaths do not affect anger responses, which are in fact exaggerated in both clinical populations (Blair, 2012; Blair and Cipolotti, 2000; Forth et al., 2003).

Moreover, the first and only experiment which tested participants answering the Knobe scenarios using magnetic resonance (Ngo et al., 2015), found that higher levels of activation of amygdala, well known for its role in processing emotions (Phelps & LeDoux, 2005), are associated with higher judgments of intentionality for the negative side effect and so, the authors concluded that this area would be involved in determining the KE. According to the authors, this relationship is mediated by judgments of *emotional reaction*[12], meaning the emotional arousal individuals felt during the reading of the scenarios, which were collected in a post-scan session. So the more the amygdala is activated, the more emotional arousal is perceived and the higher intentionality is attributed to the negative side effects. In turn, the relationship between *emotional reaction* and *intentionality* judgments for the negative side effects is further mediated by judgments of *moral blame*[13],meaning how much blame the Chairman deserves, collected afterwards (the more *amygdala* activation→ the more *emotional*

[12] How did the Chairman’s harming (or helping) the environment make you feel? [-3=Very Negative to 3=Very positive]

[13] How much blame does the Chairman deserve for helping the environment? [1=No Blame at All to 8=Extreme Blame]

*reactio*n →the more *blame*→ thus more *intentionality* attributed to the negative side effects), aligning with the Alicke's model. On the other hand, as regards positive side effects, the authors suggested a different path: judgments about positive side effects are predicted by judgments of *credit*[14], meaning how much credit the chairman deserves, which in turn are predicted by judgments about *statistical normativity*[15], meaning how common certain actions are perceived to be by the general population. In this case, the tendency to withhold credit, because the agent expresses an indifferent attitude towards something good (not following the statistical norms), decreases intentionality attributed (the lower assessment of *statistical normativity* → the lower *credit*→ the lower intentionality ratings for the positive conditions).

1.2 Alexithymic traits of personality and deficit of processing emotions

Until now, the "emotional hypothesis" of the *KE* has been explored in VMPFC and psychopaths, due to their emotional processing deficits, whereas no results have been collected in another population which also presents a wide range of emotional difficulties: individuals with alexithymic traits of personality.

Alexithymia refers to a personality trait characterized by difficulties in recognizing, describing, and verbalizing one's feelings, difficulties in distinguishing feelings from bodily sensations of emotional arousal, a paucity of imaginal capacities and an externally oriented cognitive style (Taylor, 2000). This trait of personality is of particular interest to the present investigation since prior studies showed a variety of deficits pertaining to the emotional context, especially in relation to negative emotions. For instance, emotional deficits have been described in terms of: (i) *emotional word processing*: poor processing of speech prosody with emotional content (Goerlich-Dobre et al. 2013); as well as in the recall of emotional words (Luminet et al., 2006); and in understanding others' emotions (Moriguchi et al. 2006; Swart et

[14] How much credit does the Chairman deserve for helping the environment? [1=No Credit at All to 8=Extreme Credit]

[15] About how many people out of 100 in the general population would have harmed (helped) the environment under these circumstances? [0 to 100]

al. 2009); (ii) *emotional perception*: lower accuracy during identification of facial expressions (Grynberg et al., 2012; Parker et al., 1993) and lower rating of salience concerning fearful faces (Prkachin et al., 2009); (iii) *emotional embodiment*: in particular, a diminished response to stimuli of fear was found, suggesting a hypo-functioning of amygdala in alexithymia (Scarpazza, Ph.D. dissertation, 2015); and (iv) *empathy abilities*: less distress experienced in the face of others' suffering and lower motivation to act altruistically to relieve another's distress (FeldmanHall et al., 2012) were found in these individuals.

As regards this latter ability, recent studies (Patil et al., 2014, 2016) have shown a close relationship between alexithymia and empathy in the moral domain. These studies indicated that alexithymic traits are associated with an increase of utilitarian tendencies in personal moral dilemmas (sacrificing, with personal force, one individual for the good of many) and this was due to a reduced empathic concern for the victim exhibited by these individuals. Indeed, at the opposite, according to the "empathic blame hypothesis" (Patil et al., 2017) a greater activation of the empathy network (for victims) on the part of those who judge, predicts higher condemnation of harmful acts, even to unintentional harmdoers, due to their empathy for the suffering of the victim.

Despite this close relationship however, empathy abilities have not been considered in the present investigation, since in Study 2 no relationship was found with regards to intentionality attribution.

The behavioural deficits of alexithymic individuals are associated with alterations in emotional arousal (Neumann et al. 2004; Pollatos et al. 2008; Starita et al., 2016) which seems to be due to a deficit in the early emotional reactivity to emotional stimuli even if data is weak and contradictory. Indeed, both increased and decreased physiologic reactions to emotional stimuli have been reported (Taylor & Bagby, 2004), therefore, initial studies have supported a "hyper-arousal" explanation (Stone & Nielson, 2001; Panaite & Bylsma, 2012), whereas recent investigations have found more considerable support for a "hypo-arousal" hypothesis by the discovery of lower cardiac (Pollatos et al. 2008) and electrodermal responses in these individuals (Wehmer et al. 1995; Linden et al.,1996; Franz, Schaefer, & Schneider, 2003; Neumann et al., 2004; Bermond et al., 2010a; Pollatos et al., 2008, 2011). If so, this reduced arousal might produce an ambiguous emotional context which is highly difficult to understand and interpret. Furthermore, recent examinations found a decreased activation of the amygdala associated with this diminished emotional response (e.g. van der Velde et al., 2013; Reker et al., 2010), especially in the processing of negative stimuli, suggesting a decreased attention to such stimuli in these individuals. The amygdala is widely known for its role in the detection of emotional stimuli and the redirection of attention toward these stimuli (Hodsoll et al., 2011). Consequently, when this detector is less activated, attention towards emotional or novel stimuli declines (Anderson and Phelps, 2001; Jacobs et al., 2012), as appears to be the case with alexithymic individuals. In particular, a reduced left amygdala volume has been linked to higher levels of alexithymia (e.g., Goerlich-Dobre et al., 2015; Wingbermühle et al., 2012; Grabe et al., 2014) and it has been suggested to be the basis of some of their typical difficulties in identifying feelings (Moriguchi and Komaki, 2013; Wingbermühle et al., 2012).

1.3 The present study

Due to contradictory results obtained in the literature about the role of emotions in determining the increase of intentionality judgments, in particular, towards the negative side effect of the *Knobe scenario*, the aim of the present study was to provide new evidence about this relationship. This was achieved by exploring the *KE* within a group of individuals who exhibit deficits in dealing with emotions, *alexithymic* individuals, compared to a group of non-alexithymic individuals.

Alexithymic individuals are of particular interest since they exhibit a wide range of emotional processing deficits, likely partly due to a reduced amygdala activation, which has also been found to be involved in the *KE* (Ngo et al., 2015).

Moreover, in the experiment, both subjective (*judgments of subjective arousal while reading scenarios*) and objective (*skin conductance level, hereafter referred to as SCL*) measures of arousal were collected, compared to previous investigations of the *KE* in individuals who exhibit emotional processing deficits, which only assessed the intentionality judgments.

The first hypothesis of the present study predicted reduced judgments of *subjective arousal*, and consequently lower *intentionality judgments* towards negative side effects (resulting in a reduced *KE*), which could be a result of a diminished attention towards negative emotional stimuli exhibited by alexithymic individuals.

It is important to note that previous studies (Ihme et al 2014; van de Velde et al., 2013) pointed out that the emotional processing difficulties of alexithymic individuals are particularly evident under limited presentation times of emotional stimuli, with more ambiguous emotional stimuli, or during the complex decision-making process required for moral dilemmas, whereas when faced with increased presentation time or easier emotional processing tasks, a similar performance to low alexithymic individuals is achieved. Indeed, while performing easier emotional processing tasks (for instance more salient stimuli) deficits in the automatic redirection of attention, due to defective amygdala in these individuals, could be compensated by the involvement of high-level mechanisms, such as a greater activation of the dorsal anterior cingulate cortex involved in the top-down redirection of attention towards the emotional stimuli (van der Velde et al., 2013). Whereas, this compensation mechanism might fail when task difficulty increases (i.e. with more ambiguous stimuli). In order to take these differences in emotional processing into account, participants were administered with two kinds of emotional stimuli: the classic scenarios used by Knobe represented the more complex emotional stimuli; and scenarios characterized by a greater emotional salience represented the easier emotional stimuli, since being more salient they were considered more able to elicit emotions in these individuals, by the takeover of the high-level compensatory mechanisms mentioned above.

Therefore, the second hypothesis, based on these studies (Ihme et al 2014; van de Velde et al., 2013), predicted no differences between groups in the high salience condition, compared to the standard one.

Finally, as regards the SCL measure, according to the recent hypo-arousal theory (e.g. Linden et al.,1996; Neumann et al., 2004; Pollatos et al., 2011) a similar pattern of responses was expected, with differences in the standard condition and no longer in the high salience condition.

2. Methods

2.1 Participants

Two-hundred university students completed the 20-item Toronto Alexithymia Scale (TAS-20) (Bagby, Parker, & Taylor, 1994; Bressi et al., 1996). Depending on the score, students were classified as alexithymic (A- TAS-20≥61) or non-alexithimic (NA- TAS-20≤36). Individuals from the two groups were randomly contacted and asked to participate in the study. Once in the laboratory, participants' classification was confirmed with the alexithymia module of the structured interview for the Diagnostic Criteria for Psychosomatic Research (DCPR) (Porcelli & Sonino, 2007; Mangelli et al. 2006; Porcelli & Rafanelli, 2010), previously used in alexithymia research (Grandi et al. 2011).

Due to the high co-occurrence of alexithymia and depression, participants also completed the Beck Depression Inventory II (Beck et al., 1996) and were excluded in case their score was higher than the moderate/severe depression cut-off (n= 8).

The final sample included in the analysis consisted of 44 participants (17 males, 27 females; age *M* = 21.64, *SD*= 2.71 years old) divided in two groups: 22 LA participants (14 females; TAS-20 *M* =32.00, *SD* =2.15; age *M* =21.90, *SD* =3.14 years old) and 22 HA participants (13 females; TAS-20 *M* =64.95, *SD* =2.32; age *M* =21.41, *SD* =2.30 years old). A priori target for sample size was based on similar studies' samples, employing electrodermal activity recording (e.g. Cecchetto et al., 2017).

The study was approved by the Ethics Committee of the University of Bologna. All participants gave informed written consent to participation after being informed about the procedure of the study.

2.2. Materials and procedure

2.2.1 Stimuli

The task included twenty scenarios drawn from Ngo et al. (2015) (even if the original database was the same, the scenarios employed were not the same of Study 3, for more details see Appendix B) which were modeled after the Knobe original vignette. The English version was translated before into Italian by an Italian native speaker, proficient in English. Then, the translated dilemmas were presented to an English native speaker proficient in Italian for the back-translation to English (Brislin, 1970).

Both valence and saliency of scenarios were modulated: ten scenarios involved a negative side effect and ten involved a positive one and, in turn, in each group five scenarios were characterized by a standard emotional salience and five were characterized by a high emotional salience. In the first case, like in the classic Knobe scenario, the consequences could involve the environment or people's relationships, whereas in the second case the consequences were more salient since could involve people's life and health. An example of one scenario characterized by a standard emotional salience and one characterized by a higher emotional salience is reported in Table 9. The other scenarios are reported in the Appendix B.

Condition	**Negative side effect**	**Positive side effect**
STANDARD SALIENCE SCENARIO	Jenny spread weed killer to protect her crops. Jenny did not care at all about the effect this would have on Susie-Ann's crops. Jenny knew her pesticide would harm Susie-Ann's crops.	Stanley spread anti-fungals to protect his crops. Stanley did not care at all about the effect this would have on Billy-Bob's crops. Stanley knew his anti-fungals would protect Billy-Bob's crops.
Judgment	*Did Jenny intentionally harm her neighbor's crops?*	*Did Stanley intentionally protect Billy-Bob's crops?*
HIGH SALIENCE SCENARIO	Bill released a gadget to make a deadline. Bill did not care at all about the effect the gadget would have on babies. Bill knew his gadget would kill babies.	Robyn released an invention to make a deadline. Robyn did not care at all about the effect the invention would have on toddlers. Robyn knew her invention would help toddlers.
Judgment	*Did Bill intentionally cause the death of babies?*	*Did Robyn intentionally help toddlers?*
Answer	0 = unintentional - 10 = completely intentional	0 = unintentional - 10 = completely intentional

Table 9. Example of scenarios characterized by a standard salience or a higher salience

Each scenario was followed by two questions about the intentionality of the side effect (e.g. *Did Jenny intentionally harm her neighbour's crops*?) and the subjective level of arousal experienced while reading the scenario (*How strongly did you emotionally react when reading the scenario*?), which had to be rated on a scale of 0 (not at all) to 10 (completely). The scale of response was reduced, in the present experiment, to 10 points, due to the fact that the experiment was programmed and run using E-Prime (Schneider, Eschman, & Zuccolotto, 2002) and stimuli were presented on a high-resolution computer monitor.

In each trial, the scenario appeared on the screen for 25 seconds followed by another screen, showing the first question, and by a third, showing the second question (Figure 6). The questions remained on the screen until participants provided an answer. After the second question, a blank screen was displayed for 8 seconds. The order of scenarios was counterbalanced across participants, whereas order of questions was fixed.

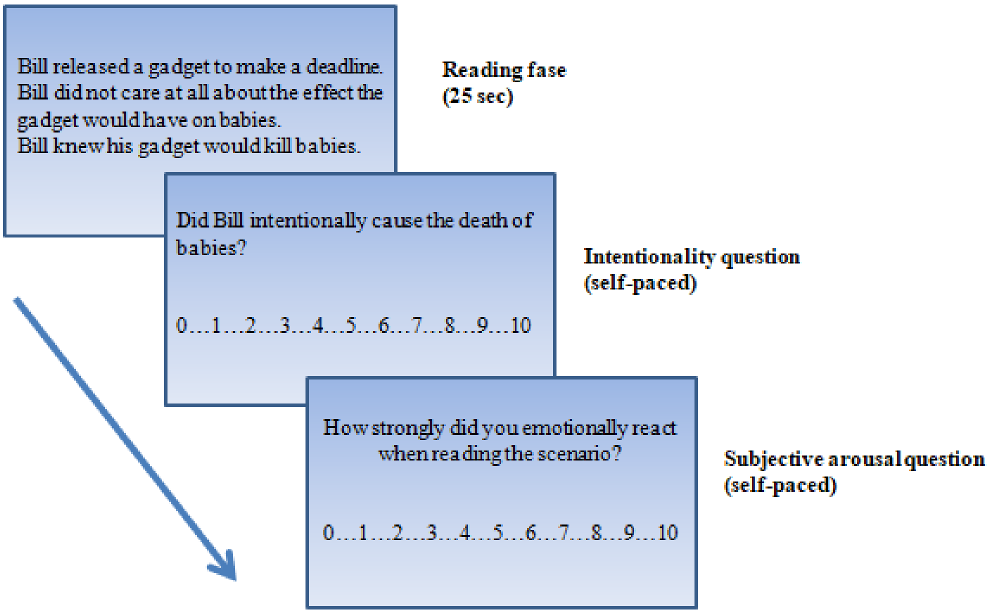
Bill released a gadget to make a deadline.
Bill did not care at all about the effect the gadget would have on babies.
Bill knew his gadget would kill babies.
Reading fase
(25 sec)
Did Bill intentionally cause the death of babies?
0…1…2…3…4…5…6…7…8…9…10
Intentionality question
(self-paced)
How strongly did you emotionally react when reading the scenario?
0…1…2…3…4…5…6…7…8…9…10
Subjective arousal question
(self-paced)

Figure 6. Example of the experimental procedure

2.2.2. Skin conductance level acquisition and processing

Skin Conductance Level (SCL) was collected as measure of emotional arousal. SCL was recorded through two Ag/AgCl electrodes (TSD203 Model; Biopac Systems, USA), filled with isotonic hyposaturated conductant attached to the distal phalanges of the second and third finger of participants' left hand and held with Velcro straps. So participants were able to answer to the questions with their dominant hand. The SCL signal was continuously recorded at 200 Hz and amplified using a DC amplifier (Biopac GSR100; Biopac Systems, USA) with 5 μ S/V gain factor and 10 Hz low pass filter to remove high-frequency noise. The analogue signal was digitalized using the MP-150 digital converter (Biopac Systems, USA) and fed into AcqKnowledge 3.9 software (Biopac Systems, USA). SCL was collected continuously during the task and stored for off-line analysis on a second PC. (Figure 7).

Data from five participants were removed due to a lack of sufficient physiological responsiveness or to technical problems during the recording. SCL data were processed using the *Autonomate toolbox* (Green et al., 2013) for MATLAB (The MathWorks, Inc., USA). SCL was extracted for three time intervals of scenarios visualization (0.5-8; 8-16; 16-25 seconds). SCL was converted into microsiemens and square root transformation was conducted on raw SCL to normalize the data distribution.

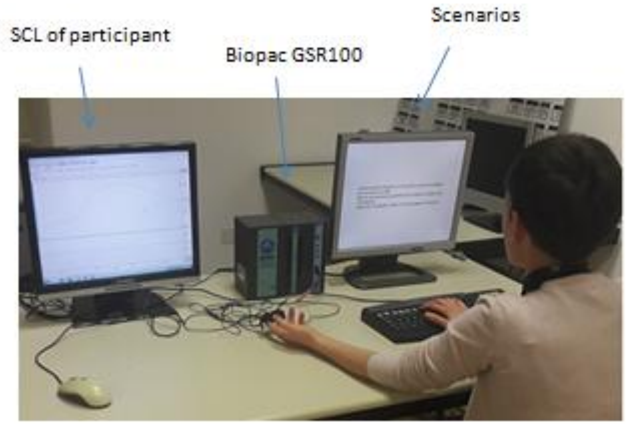

Figure 7. Example of the experimental setting.

2.2.3. Procedure

The experiment took place in a quiet room with dimmed light. Participants were seated on a chair in front of a computer monitor at ~70 cm distance. Once seated, the experiment procedure was explained and written informed consent was obtained from participants.

Each testing session began with a 5-min rest period during which the participants acclimatized to the environment. Then electrodermal activity (EDA) electrodes were attached and the correct recording of the signal was ensured as well as the responsiveness of EDA to external stimuli (e.g. loud noise made by hand clap).

During the task, participants were asked to remain silent and relaxed, and avoid movements to ensure the correct acquisition of the psychophysiological data.

The participants' task was to carefully read each scenario and then answer the questions, about intentionality of the side effects and about subjective arousal perceived while reading, by pressing one number from 0 to 10 on the computer keyboard.

Participants were told to respond as soon as they made a decision, although they knew there was no time limit to provide the answer. The experiment lasted around 45 minutes for each participant.

At the end of the experiment, a brief explanation about the experiment's goals was provided if requested. After that, the participants were thanked for their participation and dismissed.

2.2.4. Data analysis

Assumptions of normal distribution of the data were verified (Shapiro-Wilk test).

Several ANOVAs were then used to investigate differences between the two groups on each dependent measure. The average of intentionality judgments provided to the five scenarios at standard salience and the average of intentionality judgments provided to the five scenarios at high salience were considered, for both valence of side effects (as in previous studies of the present dissertation). ANOVAs were also used to investigate skin conductance differences between the two groups considering both valence, salience and the three time intervals. Post hoc analyses were conducted with Newman-Keuls test. Significance threshold was $p < 0.05$.

Finally regression analyses were performed to investigate subjective arousal and intentionality judgments' relationship in the two groups.

3. Results

3.1. Behavioural Data

In order to assess the association among demographic variables (age and gender) on the dependent variables (judgments of intentionality and judgments of subjective arousal), a series of correlation analyses were performed for each valence and salience of side effects.

Age (r=.01, *p*=.46) and gender (r=.01, *p*=.75) were not correlated to the negative judgment neither in the standard salience, nor in the high salience: age (r=.00, *p*=.57) and gender (r=.00, *p*=.92); the same was found for the positive side effect, which were not correlated to age (r=.02, *p*=.85) and gender (r=.02, *p*=.31) in the standard salience, as well as in the high salience: age (r=.01, *p*=.55) and gender (r=.01, *p*=.10).

Lack of correlation was also found for measures of subjective arousal for negative side effect in the standard salience: age (r=.00, *p*=.78) and gender (r=.00, *p*=.97) and in the high salience: age (r=.01, *p*=.53) and gender (r=.01, *p*=.93); as well as for the positive side effect in the standard salience: age (r=.00, *p*=.31) and gender (r=.00, *p*=.19) and in the high salience: age (r=.01, *p*=.67) and gender (r=.01, *p*=.11). Therefore demographic measures were not considered in the subsequent analyses.

3.1.1 Intentionality judgments

A repeated measures ANOVA (RM ANOVA) was conducted on average intentionality scores, with the side effect valence (negative, positive) and salience (standard, high) as a within-subject factor and group (A, NA) as a between-subject factor.

The ANOVA yielded a significant main effect of valence [$F_{1,42}$= 217.86; *p*<.001; Ƞ^2 =.83], indicating that intentionality judgments for the negative side effect were higher than for the positive side effect (*Knobe Effect*), as well as of salience [$F_{1,42}$=134.79; p<.001; Ƞ^2 =.76], meaning that intentionality judgments provided for the scenarios in the high salience condition were higher than the same for scenarios in the standard salience condition.

Moreover a significant three-way *valence by salience by group* interaction [$F_{1.42}$= 5.24; p=.02; Ƞ2 =.11] indicated that alexithymic individuals attributed significantly less intentionality to the negative side effects, only for the standard saliency condition, compared to non-alexithymic [NA: M=7.88, SD=0.72; A: M= 6.19, SD=0.97; p<.001]. All other comparisons were not significant (*ps*=0.29-0.91) (Figure 8).

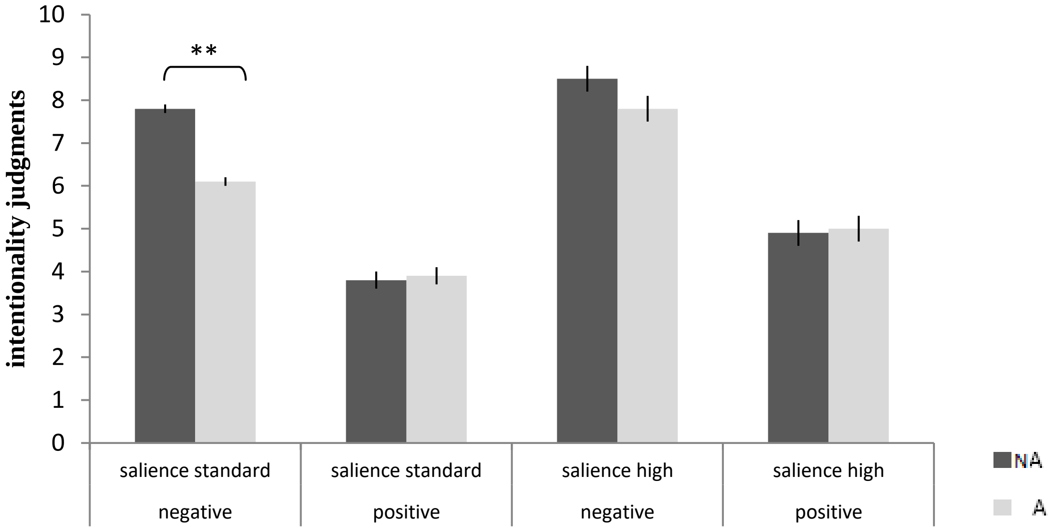

Figure 8. Means and standard errors of intentionality judgments are reported, showing a three-way interaction valence X salience X group. Compared to NA (non-alexithymic), A (alexithymic individuals) judged the negative side effect, at standard salience, as less intentionally performed. Significant differences are indicated as follows: **p<.001

3.1.2. Subjective arousal judgments

A RM ANOVA was conducted on average subjective arousal scores, with the side effect valence (negative, positive) and salience (standard, high) as a within-subject factor and group (A, NA) as a between-subject factor.

The analysis revealed a significant main effect of valence [$F_{1.42}$ =84.51; p<.001; $η^2$ =.66], indicating that participants felt more aroused reading about negative side effects compared to the positive ones; as well of salience [$F_{1.42}$=159.52; p<.001; $η^2$ =.79], meaning that side effects characterized by an higher salience were perceived more arousing that side effects characterized by a standard salience.

Moreover, a significant two-way *valence by group* interaction [$F_{1.42}$=5.45; p=.02; $η^2$ =.11] indicated that, compared to non-alexithymic, alexithymic individuals reported significantly less arousal for the negative side effects, regardless of the saliency of the scenario [NA: M=5.3, SD=1.5; A: M= 3.9, SD=1.8; p<.05]. (Figure 9)

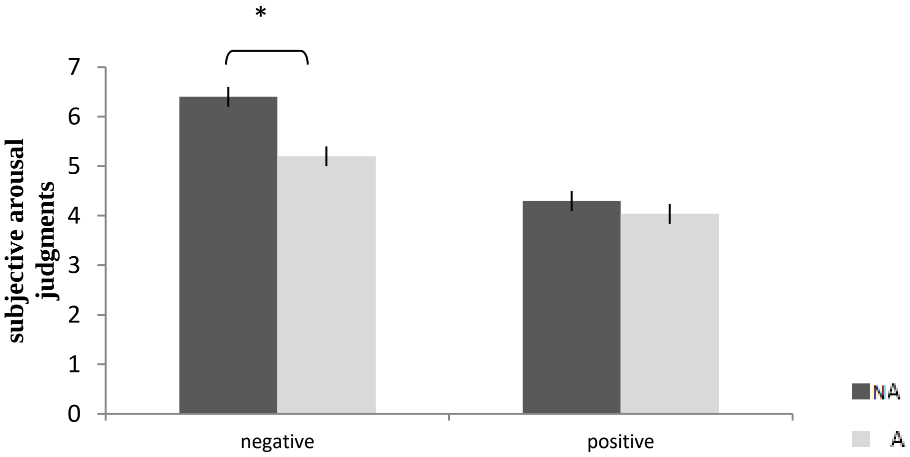

Figure 9: Means and standard errors of subjective arousal judgments are reported, showing a two-way interaction valence X group. Compared to non-alexithymic (NA), alexithymic individuals (A) reported a significantly lower subjective arousal felt when facing with the negative side effects, regardless the saliency of the scenario. Significant differences are indicated as follows: * p<.05

3.2. Psychophysiological Data

A RM ANOVA was conducted on average skin conductance level, with the side effect valence (negative, positive), salience (standard, high) and time interval (0.5- 8; 8-16; 16-25 seconds) as a within-subject factor and group (A, NA) as a between-subject factor.

The analysis revealed a significant main effect of salience [$F_{1.37}$=26.11; p<.001; η^2 =.41], indicating higher SCLs elicited from scenarios characterized by an higher salience, compared to scenarios of standard salience; as well as of group [$F_{1.37}$=14.30; p<.001; η^2 =.27], due to a lower signal exhibited from the group of alexithymic compared to non-alexithymic, regardless the salience; and a main effect of epoch [$F_{2,74}$=8.44; p<.001; η^2 =.18], suggesting an higher SCL activation elicited in the first epoch of scenario visualization, compared to the two subsequent epochs.

Moreover, a significant two-way *epoch by group* interaction [$F_{2,74}$=3.41; p=.037; η^2 =.08] indicated that, compared to alexithymic, non-alexithymic individuals exhibited a higher activation in the first epoch (0.5-8sec), than the other two epochs (8-16 sec; 16-25 sec), regardless the saliency of the scenarios (Figure 10).

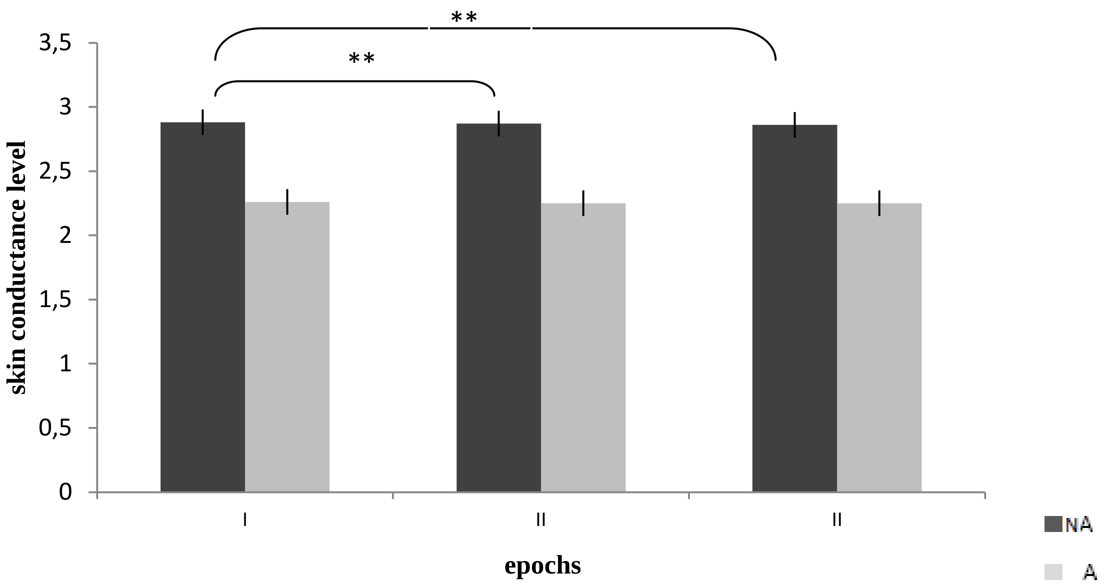

Figure 10. Means and standard errors of skin conductance level (SCL) are reported, showing a two-way interaction epoch X group. Compared to alexithymic (A), non-alexithymic individuals (NA) exhibited a greater SCL increase in the first epoch (0.5-8sec), than the other two epochs (8-16 sec; 16-25 sec), regardless the saliency of the scenarios. Significant differences are indicated as follows: ** p<.001

3.3. Judgments of subjective arousal predict intentionality judgments only for non-alexithymic individuals

In order to assess if judgments of subjective arousal predict the ascription of intentionality, for each group, two regressions were performed.

As regards NA the analysis was significant, indicating that subjective arousal perceived during reading of scenarios with a negative side effect, in the standard salience condition, $F_{1,20}$ =5.57, p<.05, ß=.46 adjusted R^2=.17, as well as in the high salience condition, $F_{1,20}$ =7.47, p<.05, ß=.52 adjusted R^2=.23 predict the intentionality judgments provided. The higher was the subjective arousal perceived, the more increased are intentionality judgments provided, regardless the emotional salience.

Instead, as regards A, judgments of subjective arousal did not predict intentionality judgments provided, neither in the standard salience condition (p=.23), nor in the high salience condition (p=.83).

4. Discussion

In the present study the role of emotions in determining the increase of intentionality attributed to the negative side effect of the Knobe scenario was explored in a group of alexithymic individuals.

According to the literature, these individuals exhibit a wide range of emotional processing deficits, pertaining to several areas such as emotional perception, word processing, and empathy abilities (e.g. Goerlich-Dobre et al. 2013; Grynberg et al., 2012; Patil & Silani, 2014). In addition, they exhibit an associated atypical emotional arousal (e.g. Neumann et al. 2004; Pollatos et al. 2008) which could explain the origin of their emotional awareness impairment. A recent review (Van der Velde et al., 2013) highlighted that decreased activity in the amygdala has been constantly found in alexithymics, which led to the hypothesis that it is responsible for the reduced attention to the emotional stimuli observed in these individuals, due to its role in the detection of emotional significance and the generation of emotional feelings (Bermond et al., 2006; Goerlich-Dobre et al., 2013; Kano and Fukudo, 2013; Larsen et al., 2003; Moriguchi and Komaki, 2013; Taylor and Bagby, 2004; Wingbermühle et al., 2012).

The same area was found to be involved in the *KE,* and it has been considered responsible for the increase in intentionality judgments when evaluating negative side effects (Ngo et al., 2015). Thus the performance of these individuals who are characterized by atypical emotional reactions, when facing the Knobe scenarios could allow us to understand whether these emotional reactions are necessary for the increase of judgments.

First of all, the results of the present study indicated the presence of the *KE* and so, in both groups examined, the negative side effects were considered to be produced with a significant higher intent, compared to the positive side effects. This result is in line with a great amount of previous research (for a review see Feltz, 2007).

The last study (Ngo et al.,2015) which supported the role of emotions in the *KE* found that only the *emotional arousal* experienced when facing negative side effects allows one to predict the subsequent intentionality attributed, whereas no positive relationship was found for side effect characterized by a positive or neutral valence (greater emotional arousal→ higher intentionality ascribed to the negative effects). According to the authors, amygdala is primarily responsible for the increase of judgments, since the more *amygdala* is activated, the higher *subjective arousal* was perceived, thus higher intentionality was assigned. In our study a further step was added, exploring the evaluations of participants when facing scenarios characterized by a greater emotional salience, compared to the standard emotional scenarios (of the classic Knobe scenario). This was implemented, in particular, because alexithymic individuals exhibit their emotional processing deficits especially towards more complex emotional stimuli, whereas they produced a similar performance to non-alexithymics when facing easier emotional tasks (Ihme et al 2014; van de Velde et al., 2013).

Therefore, the second step of the present study was to verify the influence of the salience of side effects: as expected, side effects characterized by a higher emotional content were perceived as more intentionally performed, compared to the classic side effects. Thus, the more emotionally salient the consequences in the scenarios are, the more they are perceived as intentionally achieved. Interestingly, this phenomenon also occurred for positive consequences, showing that by increasing the emotional salience of *positive effects,* they are also perceived as more intentional. This result is consistent with a prior study of Nadelhoffer

(2004a), which showed that if the *positive effect* is made more praiseworthy, it could be considered intentional as well. According to these results, the emotional salience seems to modulate intentionality attribution at large, regardless of valence, despite the fact that a different degree of salience seems to be necessary for the two valences (positive and negative) to modulate intentionality: for negative side effects a standard salience is enough, whereas for positive side effects, the salience must be higher in order to influence the intentionality attribution.

As regards the first hypothesis, which predicted lower judgments of intentionality exhibited by alexithymic individuals, the present study confirmed that, although present, the *KE* is less pronounced because alexithymic individuals perceived the negative side effects, at the standard salience, as produced with less intent compared to non-alexithymic individuals. This result is consistent with the hypothesis of a reduced amygdala activity in alexithymics (Van der Velde et al., 2013) which might lead to a decreased attention towards negative stimuli, perceiving them as less arousing. This is also consistent with the results obtained in our study, from both subjective (*judgments of subjective arousal*) and objective (*SCL*) measures of arousal, which indicated a reduced activation in alexithymic participants, compared to non-alexithymic ones, when facing the negative side effects characterized by a standard salience (like the classic Knobe scenario). Moreover, a perceived greater *subjective arousal* only led to judgments of higher intentionality for the negative side effect, at the standard salience, in the group of non-alexithymic (as in Ngo et al., 2015), whereas the same judgments were not predictive in the group of alexithymic, who do not seem to be informed by their subjective awareness.

This result would partly provide support for the “emotional hypothesis” since the ability of perceiving emotions (at least as subjectively perceived) while evaluating side effects, seems to influence the attribution of intentionality to those side effects. Moreover, it is consistent with studies which posited that when faced with more ambiguous or complex emotional stimuli, such as classic Knobe scenarios, the defective automatic attention of alexithymic individuals does not allow them to appreciate that stimuli, which tend to be less detected, compared to what occurs in the case of non-alexithymics (Grynberg et al., 2012; Prkachin et al., 2009; Parker et al., 2005; Eichmann et al., 2008; Ihme et al., 2014).

On the other hand, in line with hypothesis two, when facing stimuli characterized by a higher emotional content, the difference concerning intentionality attribution, between groups, disappears and alexithymic individuals produce a similar performance to non-alexithymics. This suggests that the higher salient stimuli, being easier, might activate the involvement of compensatory mechanisms, other than amygdala (van der Velde et al., 2013), which allow the redirection of top-down attention towards the stimuli, even if the *emotional arousal* perceived by alexithymics (as indicated by both *SCL* and *subjective measures* of arousal) remains lower than that of non- alexithymics.

As regards *subjective arousal judgments*, negative side effects were perceived as more arousing compared to the positive ones, regardless of the salience; and they were perceived to be more salient in the high salient condition compared to the standard one, in both groups of participants. Therefore, alexithymic individuals also perceived an *increase in subjective arousal,* associated with an increase in *SCL*, when the emotional content of scenarios is higher, however, contrary to our second hypothesis, although increased, their emotional arousal was still lower than that of non- alexithymics. In fact, alexithymics still experienced a lower *subjective arousal* in the high salience condition, and they did not exhibit any difference among the three epochs of *SCL*, compared to non-alexithymics who showed a higher increase of *SCL* within the first 8 seconds of scenarios' reading.

As regards measures of arousal, they were both found to be reduced in our sample. In recent literature, a lot of studies have found a decoupling between the two measures: *physiological response* to emotional stimuli of alexithymic individuals was found to be both equivalent (Bausch et al., 2011; Stone et al., 2001) or decreased (Bermond et al., 2010; Neumann et al.,

2004) compared to a control group; the same was true for *subjective reports* of emotional experience, which were the same (Franz et al., 2003), increased (Pollatos et al., 2008; Eastabrook et al., 2013) or decreased (Stone et al., 2001) compared to a control group. Nevertheless, the dampening measures of arousal found are consistent with the "hypo-arousal" hypothesis of alexithymia (Linden et al.,1996; Franz, Schaefer, & Schneider, 2003; Bermond et al., 2010; Pollatos et al., 2011; Neumann et al., 2004), and could explain the emotional awareness difficulties of these individuals as due to their reduced psychophysiological emotional experience. An alternative explanation could consider the task administered to participants, which is characterized by very different and more complex content ("establish the intentionality of a complex action which is a side effect") compared to the classic tasks used to explore emotional abilities in alexithymia. Indeed this result is consistent with results obtained in a task with a comparable level of difficulty, such as providing moral judgments (Cecchetto et al., 2017). Similarly, the authors found lower SCR in individuals with high alexithymia compared to those with lower alexithymia scores.

In summary, individual differences in dealing with emotions seem to strongly influence the attribution of intentionality: despite negative side effects being perceived as more arousing and intentional by both groups compared to the positive ones, reduced *SCL activation* and *subjective emotional arousal* towards negative stimuli in alexithymic individuals led to the perception that the action was less intentional, compared to the responses of non-alexithymics. However, this only occurred with stimuli characterized by a standard emotional content, since they are likely to be more complex and ambiguous for these individuals.

On the other hand, when the emotional stimuli were highly salient, despite a lower arousal in alexithymic compared to non-alexithymic individuals, an increase in all measures occurred (*SCL, subjective and intentionality judgments*) and there were no longer any differences between groups in the intentionality ascription. As regards intentionality judgments, the *emotional salience* characterizing the second kind of scenarios (those which involve people's lives and health) seemed to operate as a “bait” to activate alternative strategies; for instance when stimuli are so salient that alternative cognitive mechanisms could take over, rendering them impossible to ignore.

So, the present investigation revealed new evidence about the “emotional hypothesis”, which was found to only partially explain the *KE*. Other mechanisms take over and redirect the attention towards the more salient stimuli. These mechanisms, as suggested by Diaz et al (2017) and previously by Adams & Steadman (2004a), could be the “motivation to blame the protagonist for his action”.

Moreover, the present study (as previously carried out with moral dilemmas) enlarged the investigation to different and more complex emotional stimuli such as the description of events, which involved decision-making about intentionality, compared to classic emotional tasks generally used to investigate alexithymic abilities.

Among the range of emotions, the specific emotion of *anger*, which is considered the one more associated to judgments of blame (e.g. Lerner et al., 1998; Goldberg et al., 1999), instead has been recently found not to be related to intentionality attribution (Diaz et al., 2017)[16]. In our study, the relationship between anger and intentionality attribution was not explored, but cannot be excluded, since a reduced intentionality attribution provided by alexithymic individuals (in the standard scenarios) is consistent with a previous study (Neumann et al., 2004), which found that greater levels of alexithymia are related to attenuated autonomic reactivity to angry events. However, evidence about this relationship is still limited.

Finally, it is important to note that a recent study (Slavny & Moore et al., 2017) found a positive relationship between empathy and intentionality (higher empathy ability is associated

[16] However it is important to underline that the authors, except for the classic Knobe scenario, used two other scenarios (the Planner from Phelan and Sarkissian, 2008 and the Modified Lieutenant from Phelan and Sarkissian, 2009) which are characterized by a different structure compared to the scenarios used in the present dissertation: the valence of the side effect is negative, but the agent’s main goal is positive and not with a non-defined valence.

with higher intentionality attributed), even if the task used was different to ours, since it required the attribution of intentionality to short sentences describing ambiguous main actions and not side effects.

It is possible that the presence of reduced empathy abilities could also have led to reduced compassion and concern towards the victims of scenarios and less blame towards the agent (empathic blame hypothesis- Patil et al., 2017). However Study 2 of the present dissertation failed to find a direct relationship between empathy and intentionality. Therefore we suggested the lack of direct relationship between the two, which could be mediated by emotional arousal. Indeed, Patil & Silani (2014) also mentioned an alternative hypothesis to their results, according to which people who suffer from persistent confusion about their intense emotional experiences (such as alexithymic individuals), could develop a down-regulating negative affect, associated with reduced empathic concern for the victim, which leads them to endorse more utilitarian judgments.

Therefore, further studies could consider how the association between lower “emotional activation” and lower “empathy abilities” could contribute to reduced blame and the consequent reduced intentionality judgments (emotional activation + empathy → blame→ intentionality); as well as investigating the alternative hypothesis that “empathy abilities” influence “blame attribution” and “emotional reactions” which are, in turn, linked to intentionality attribution (empathy→ blame + emotional activation→ intentionality), thus highlighting a closer relationship between “emotional reaction” and intentionality and an indirect connection between “empathy abilities” and intentionality (which is mediated by “blame and emotional activation”).

However the latter suggestions remain speculations as they are currently untested. Further research on this topic, considering all elements (e.g. empathy abilities, emotional abilities) and their interaction/mediation effects could help to clarify this framework.

Acknowledgment: Thank to Dr. Marzocchi M. for helping me in implementing the experiment on E-Prime

Chapter 3. General discussion

3.1 The role of individual differences on intentionality decision making

There is no doubt that the ability to attribute meaning to others' actions is fundamental in everyday life and sets the course of all social interactions among humans: this ability is called *attribution of intentionality*.

As described in the Introduction, this ability develops early within human abilities: it starts with allowing children to distinguish between goal-directed human actions and other events (e.g., Sommerville & Woodward, 2005; Wellman & Phillips, 2001) until adulthood, when it allows adults to understand one's own and others' behaviour in terms of underlying mental causes (Malle, 1999). "*Such judgments are so deeply engrained in human cognition that we might count intentionality alongside space, time, and causality as one of the fundamental categories with which the mind makes sense of the world*" (Malle, 2006).

Therefore, this capability is a requirement for coordinated social interactions in everyday life, and plays a crucial role in the legal field, where criminal responsibility weaves a close link with judgments of intentionality. Indeed, the presence of intent, on the part of the agent, represents the crucial element implying a responsibility for that act, whereas if this element is lacking, the consequent degree of responsibility is reduced (Malle, 2006).

Our assessment of intentionality seems to be effortless, and there is substantial agreement among individuals about which kind of actions should be considered intentional: in general, actions generated from desires and beliefs are considered to be intentionally produced, whereas causes external to the agent tend to be used to explain the unintentional outcomes. Yet, despite the apparent ease with which we make these decisions, it has been suggested that a greater tendency might exist among people to interpret actions as intentional, compared to the opposite, due to an adaptive heuristic which allows faster answers to the social stimuli around us. Thus, intentional explanations of actions would be automatically triggered and represent our default for explaining others' behaviour (Rosset, 2008). The present dissertation has attempted to explore why people tend to express this "over-attribution" towards side effects, and in particular when the valence is negative (Knobe, 2003). Starting from the first paper on the *KE* (Knobe, 2003), a wide volume of research flourished in several contexts including philosophy, psychology, and neuroscience, in the attempt to explain the underlying mechanism. As largely explained above, several hypotheses have been proposed, ranging from the influence of prior morality evaluations which would, in turn, influence intentionality ascription (Knobe, 2006); to the influence of emotions and blame elicited from the negative consequences while reading the scenarios (Malle and Nelson, 2003) and many more; all of these were attempted to find a "unitary mechanism" able to explain the *KE,* and intentionality attribution at large.

Instead, the possibility that individuals might express different intuitions about important topics, such as intentionality, according to their individual characteristics has been under-estimated for a long time. Even though, in many other domains the folk intuitions of individuals have been found to be fragmented according to their individual characteristics (Feltz & Cokely, 2008). Therefore, it is crucial to take those individual differences into account, exploring what they can reveal about the mechanisms of intentionality ascription.

Previous studies (Mele & Cushman, 2007) exploring the role of individual differences on intentionality ascription in the *KE* showed that, according to these differences some individuals tend to consider "belief" information as a sufficient condition to attribute intents to the agent, whereas others consider a "desire" for the action, on the part of the agent, as necessary in order to explain the behaviour as intentional; or even, the higher presence of extraversion traits of personality led to give more importance to the negative side effects due to the association of this trait with looser regulation of affective reactions (Cokely and Feltz, 2009); or even our individual capability to imagine how the agent could have acted differently (ability of *counterfactual reasoning*) led to consider new elements, such the alternatives that the agent could have undertaken to avoid the situation, and this, in turn, influence our attribution of intentionality (Byrne et al., 2013).

Starting from the individual ability to reason counterfactually, we explored some individual characteristics which had a significant influence on intentionality attribution: the ability of counterfactual reasoning, the ability of theory of mind and the ability to process emotions.

Counterfactual reasoning has previously been shown to influence intentionality ascription, since its spontaneous production leads individuals to focus on different elements when analyzing a social situation (Byrne et al., 2013). However, previous studies (e.g. Byrne et al., 2013; Pellizzoni et al., 2010) did not distinguish between the effect of upward and downward counterfactuals.

The results of Study 1 investigated this distinction on intentionality attribution for the first time, highlighting that when people produce downward counterfactuals, they consequently judge the agent's action as less intentional; whereas when producing upward counterfactuals, people tend to perceive the agent's behaviour as more intentional. This occurred because in the case of downward counterfactuals, people compare the circumstances in which an event happened with the circumstances in which it has not happened, and which the participants imagine could have been more dramatic (e.g. Tasso, 1999). On the other hand, in the upward case, people imagine how the event could have been better, and they tend to believe that a causal link exists between the current actions taken by the defendant before the occurred event and the negative outcomes. Indeed imagining worse scenarios leads people to consider the negative consequences as less foreseeable for the agent, and for this reason they attribute less intention for such outcomes to him/her. Conversely, when people think about better scenarios, they focus on what the defendant could have done as an alternative, and they are more prone to think that negative results could have been avoided, thus attributing greater intentionality to him/her. This indicates that when generating one kind of counterfactual or the other, individuals focused on different elements: downward ones drive the observer to focalize on *external circumstances,* perceiving the agent as less able to foresee the consequences; whereas upward ones lead to a focus *on the agent himself* and on his/her ability to foresee the negative consequences of his/her behaviour. These results are of crucial importance since several studies have shown significant individual differences among people in their tendency to undertake one or the other kind of counterfactual according to their personality characteristics, even if they have experienced similar outcomes (e.g., Kahneman & Miller, 1986). These characteristics influence both the activation and type of counterfactual undertaken which, in turn, influence the focus of attention on different elements in social situations (*external consequences* or the *agent himself*), and lead to an opposite attribution of intentionality.

A focus of attention on different elements pertaining to the social situation being evaluated also emerged in Study 2. According to their individual ability of *Theory of Mind,* individuals focused on different elements which led them to an opposite attribution of intentionality. The Study analyzed its influence on intentionality attribution, whereas previously the role of Theory of Mind ability has been usually considered, for the most part, regarding individuals suffering from autism spectrum disorder or the domain of morality (e.g. Young et al., 2013; Fu et al., 2014). Results of the study showed that thanks to higher abilities in understanding others' mental states, in particular thanks to cognitive Theory of Mind ability, individuals provided decreased judgments of intentionality to the negative side effect, whereas individuals with lower Theory of Mind abilities provided higher judgments. Indeed, in the first situation it was revealed that higher abilities focus the attention, of those who judge, on the information about the "intent of the agent" to a much greater extent and for this reason the side effect was perceived not to be linked to the agent's desires, and they lowered their intentionality judgments. On the other hand, the group with lower cognitive Theory of Mind abilities focused on side effects to determine their higher judgments. These results showed us again that according to our capability to focus and be attracted more or less to one piece of information, as opposed to another, the consequent attribution of intentionality we made can be radically changed and this, in turn, could influence the associated attribution of responsibility and the severity of punishment (as found in Study 1).

The same redirection of attention to specific elements, according to our individual abilities was found again in individuals with high levels of autistic traits. These individuals, as recently highlighted, exhibit atypical Thoery of Mind abilities (Gökçen et al., 2014, 2016; Lockwood et al., 2013) concerning the ability of decoding the mental states of others (*Theory of Mind abilities*; Abell et al., 2000 Baron-Cohen et al., 2001a; Lockwood et al. 2013), compared to spared empathy abilities (Jones et al. 2010; Schwenck et al. 2012).

The results of Study 3 add new evidence, clarifying that these difficulties pertaining to both cognitive and affective subcomponents of Theory of Mind ability. These poor abilities to redirect attention towards the information about intent mediate the relationship between autistic traits and attribution of intentionality, leading attributions of these individuals to be more based on information about side effects. Moreover, both the lack of a robust Theory of Mind, as well as of efficient executive resources (poorer cognitive flexibility and repetitive behavior are typically belonging to the autistic trait pattern), hinder the overriding response driven by the emotionally salient side effects, leading these individuals to adopt a "generalized approach to the judgments", and increasing judgments towards both positive and negative stimuli.

The opposite pattern occurred (at least as regards the classic situation of Knobe) for alexithymic individuals, who due to their deficit in emotional processing (e.g., Roedema &Simons, 1999) were less attracted to the salient information of the side effect.

At the opposite, in agreement with the latest models on ASD (e.g. Bird & Cook, 2013; Rogers et al., 2007; Smith, 2009; Lockwood et al., 2013), individuals with autistic traits of personality can likely react to emotionally salient information. The results of Study 4 led new evidence about the debated role of emotions ("the blame and affective bias hypothesis") in determining the increase of intentionality judgments for negative side effects, proving that this hypothesis is only partially supported. In line with this hypothesis, alexithymic individuals were less involved in the situation, as shown by their lower exhibited arousal and attribution of less intentionality to the side effects. This led to a reduced *KE*, which was not completely eliminated. Instead, when the salience of the side effects is higher, an increase of emotional arousal also occurred in alexithymic individuals, however it still remained lower than for non-alexithymics. Interestingly, the increase of salience influenced the judgments of intentionality on both negative and positive side effects, showing that even the latter may be considered intentional if their consequences are stronger. Nevertheless, the necessary degree of emotional salience seems to be different according to the valence of the side effect: for the negative consequences the salience of the classic scenarios is enough; whereas the emotional salience should be higher to increase the judgments towards the positive side effects (see also Nadelhoffer, 2004). So, even if the arousal reported from alexithymic individuals was still lower compared to non-alexithymic, other cognitive mechanisms seem to take over (e.g. van der Velde et al., 2013) and made it impossible to ignore the stimuli as too emotionally salient.

What overall emerged from the present experiments? First of all, results from Experiment 1 to 4 showed that the *KE* has a strong effect. Even though in our samples characterized by specific individual characteristics the intentionality judgments were found to be increased or decreased according to these features, the *KE* occurred in the majority of participants in any case. Despite representing a specific situation, the *KE* allowed to show that, in some circumstances, common people can be more attracted to the outcome information compared to the information about intent. However, the individual characteristics examined above showed to modulate this tendency towards different directions.

The second main result regards the role played by the individual differences in the ascription of intentionality. Taken together all four studies conducted showed coherently that, according to their individual difference, people are led to focus their attention, or are spontaneously attracted, to different elements when analyzing the given social situation, in order to provide judgments about intentionality. This mechanism was already shown by few previous studies on individual differences in the literature of *KE* (e.g. Mele & Cushman, 2007), which indicated that people attribute more or less importance to *belief* compare to *desire* according to those differences.

This also emerged in the present studies: according to the kind of counterfactuals elicited people focused more on *external causes* (downward) or on *the agent* (upward); the same occurred according to their *Theory of Mind* abilities which led individuals to be more attracted by side effect information compared to the intents of the protagonist, and lastly, according to their individual ability to deal with *emotions*, individuals can be more or less attracted by salient side effect information.

This focus of attention on different elements would, in turn, influence the understanding of the social situation, and then the corresponding attribution of intentionality and responsibility for the event to the agent, as well as the punishment ascribed. The crucial aspect to consider is that these abilities could lead to opposite attributions, despite the same given circumstances.

This leads to the consideration that the current models of intentionality are somehow incomplete and need to be further elaborated. For instance, Malle and Knobe's model (1997) informed us about the necessary components required for an action to be considered intentionally performed. However, until now nothing has been said about which factors could, in turn, influence these components. In light of this, the studies of the present dissertation confirmed what was already initially outlined in Mele & Cushman (2007) and Byrne et al. (2013) studies: our individual abilities determine which of the components of the model are given more attention. For instance, according to their individual abilities, individuals with autistics traits focused more on *outcome* information, compared to the *belief* information, whereas alexithymic individuals showed the opposed pattern, being less attracted to the *outcome* information.

Therefore, we would suggest that the classic models should also consider what is above the classic intentionality components, because the present results indicated that individual abilities determine the focus of our attention to one or the other component, and in turn the final judgments provided.

From a practical point of view, being aware of these individual differences represents a wealth of knowledge, and there are many potential applications of these findings. For instance, results of these studies could inform the legal field about the multiplicity of influences which can occur when we attempt to attribute the intentionality of an event. The present evidence underlined that this decision-making could be arbitrary and determined from external features to the trial. Jurors, as well as attorney and other legal actors could be more aware and more able to identify these factors which can influence their perception, in order to provide more objective judgments about criminal responsibility and intentionality (Malle & Nelson, 2003; Nadelhoffer, 2006).

Moreover, as regards the specific traits of personality (autistic and alexithymic) analyzed some important considerations can be made: the present results revealed the differences in understanding social situations between these two groups, leading to the idea that: in the case of individuals with autistic traits of personality, cognitive Theory of Mind must be a key process to consider when designing potential intervention programs supporting social abilities in these individuals; whereas the ability to process emotions, as well as empathic abilities, should be the target of potential intervention programs focused on alexithymic individuals, given the high incidence of both traits in the general population (Baxter et al., 2015; Bird & Cook, 2013).

Interventions which typically focus on broader aspects of social interactions (improving conversational skills, interpersonal relationships) could have advantages from first addressing more ‘basic’ cognitive and affective abilities, to targeting subsequently higher-order social competencies, such as the attribution of intentionality to an event or to a side effect.

Limitations

Our studies suffer from some limitations that must be held in consideration in future research. First of all, the choice of using written tasks could have reduced the generalizability of present results compared to decisions taken every day in real life social contexts. Notwithstanding this observation, all previous studies (for an overview see Feltz, 2007) explored this topic in the same manner and therefore we preferred not to introduce other elements (linked to the administration) which could have confused the results. At the same time, this methodological choice could have reduced our understanding of the phenomenon, especially as regards the last experiment on alexithymic traits. Not considering the difficulties associated with the attempt to adequately “translate” the Knobe situations into a different means (such as a visual one), this operation could had led to obtaining further results. Indeed previous studies in the moral domain have shown significant differences in the results obtained using a different means of administration (Patil et al., 2016): for instance moral dilemmas in virtual reality being perceived as more emotionally arousing than the ones in text, suggesting that moral decision-making in hypothetical moral dilemmas is susceptible to the contextual saliency of the presentation of these dilemmas. The same could occur in the domain of intentionality decision-making.

Another limit could be not having considered the individual characteristics examined together, for instance autistic traits and alexithymic traits of personality (Lockwood et al., 2013; Bird & Cook, 2013). Again, in the moral domain, these personality traits showed divergent roles in determining utilitarian moral judgments in adults with autism: autistic traits were associated with reduced *utilitarian bias*, due to the elevated personal distress of demanding social situations, whereas alexithymic traits were associated with an increased bias, on account of reduced empathic concern for the victim (Patil et al., 2014). Even if having considered each trait individually has allowed us to better analyze each individual contribution, further studies and new analyses considering both traits could make up for this limit.

In conclusion, in the present dissertation the relevant role played by some individual differences in determining the attribution of intentionality to an event or to a side effects was explored. According to these characteristics, individuals focused their attention on different elements while analyzing social situations, which are consequently perceived in a different manner. This process, in turn, leads to a different attribution of intentionality to the event, as well as of responsibility and length of punishment. Therefore in future model of intentionality should be crucial also consider what is above the components necessary to judge a behaviour as intentional, in terms of individual differences.

Finally, the current research supported the idea of no one single mechanism at the basis of the *Knobe Effect*, whereas it suggested that the observed judgments asymmetry could be multiply determined and influenced from some individual differences.

References

Aaron, R. V., Benson, T. L. & Park, S. (2015). Investigating the role of alexithymia on the empathic deficits found in schizotypy and autism spectrum traits. *Personality and Individual Differences*. 77, 215–220 doi : http://dx.doi.org/10.1016/j.paid.2014.12.032.

Abell, F., Happé, F. G. & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*. 15, 1–16. doi: 10.1016/S0885-2014(00)00014-9

Adams, F. (2015). The Knobe Effect and the law. *Methode*, 6, 121-135.

Adams, F., Steadman, A. (2004a). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64, 173-181. doi:10.1111/j.1467-8284.2004.00480.x

Adams, F., and Steadman, A. (2004b). Intentional actions and moral considerations: Still pragmatic. *Analysis*, 64, 264-26

Alicke M. D. (2000). Culpable control and the psychology of blame. *Psychological Bullettin.* 126 556–574. 10.1037/0033-2909.126.4.556

Alicke M. D. (2008). Blaming badly. *Journal of Cognition and Culture*. 8 179–186. 10.1163/156770908X289279

Alicke M. D., Rose D. (2012). Culpable control and causal deviance. *Social and Personality Psychology Compass* 6723–735. 10.1111/j.1751-9004.2012.00459.x

American Psychiatric Association. (2014). Ed. it. Massimo Biondi (a cura di), DSM-5. Manuale diagnostico e statistico dei disturbi mentali, Milano: Raffaello Cortina Editore. ISBN 978-88-6030-661-6

Anderson, A.K., Phelps, E.A., (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* 411, 305–309.

Ask K., Pina A. (2011). On being angry and punitive: how anger alters perception of criminal intent. *Social Psychological and Personality Science* 2, 494–499. 10.1177/1948550611398415

Bacon, A. M., Walsh, C. R., & Martin, L. (2013). Fantasy proneness and counterfactual thinking. *Personality and Individual Differences*, 54(4), 469–473.

Baez S, Couto B, Torralva T, Sposato LA, Huepe D, Montañes P, Reyes P, Matallana D, Vigliecca NS, Slachevsky A, Manes F, Ibanez A. (2014). Comparing moral judgments of patients with frontotemporal dementia and frontal stroke. *JAMA Neurology*, 71(9):1172-1176. doi:10.1001/jamaneurol.2014.347

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182. doi: http://dx.doi.org/10.1037/0022-3514.51.6.1173

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42, 241-251. doi: 10.1111/1469-7610.00715

Baron-Cohen, S., Wheelwright, S., Skinner, R. , Martin, J. & Clubley, E. (2001b). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, 31, 5. doi:10.1023/A:1005653411471

Bausch, S., Stingl, M., Hartmann, L.C., Leibing, E., Leichsenring, F., Kruse, J., Stark, R. & Leweke, F. (2011).Alexithymia and script-driven emotional imagery in healthy female subjects: no support for deficiencies in imagination. *Scandinavian Journal of Psychology*. 52**,** 179–184

Baxter, A.J., Brugha, T.S., Erskine, H.E., Scheurer, R.W., Vos, T., Scott, J.G. (2015). The epidemiology and global burden of autism spectrum disorders. *Psychological Medicine*. 45(3):601-13. doi: 10.1017/S003329171400172X.

Bermond, B., Bierman, D. J., Cladder, M. A., Moormann, P. P. & Vorst, H. C. M. (2010). The cognitive and affective alexithymia dimensions in the regulation of sympathetic responses. *International Journal of Psychophysiology*. 75**,** 227–233

Bermond, B., Vorst, H.C.M., Moormann, P. (2006). Cognitive neuropsychology of alexithymia: implications for personality typology. *Cognitive Neuropsychiatry* 11, 332–360.

Bird, G. & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry*, 3, e285, doi:10.1038/tp.2013.61

Bird, G., Silani, G., Brindley, R., White, S., Frith, U. & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain: A Journal of Neurology*, 133, 1515–1525. doi: 10.1093/brain/awq060

Bird, G., & Viding, E. (2014).The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience and Biobehavioral Reviews*, 47, 520-532. doi: 10.1016/j.neubiorev.2014.09.021

Blair, R.J. (1999). Psychophysiological responsiveness to the distress of others in children with autism. *Personality and Individual Differences*, 26, 477–485. doi: 10.1016/S0191-8869(98)00154-8

Blair, R.J. (2012). Considering Anger from a Neuroscience Perspective. WIREs *Cognitive Sciences*3: 65–74.

Blair, R.J., and Cipolotti, L. (2000). Impaired Social Response Reversal: A Case of 'Acquired Sociopathy'. *Brain* 123: 1122–41.

Branscombe, N. R., Owen S., Garstka T. A., & Coleman J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome? *Journal of Applied Social Psychology*, 26, 1042–1067.

Bressi, C., Taylor, G. J., Parker, J. D. A., Bressi, S., Brambilla, V., Aguglia, E. et al., (1996). Cross validation of the factor structure of the 20 item toronto alexithymia scale: An italian multicenter study. *Journal of Psychosomatic Research*, 41, 551-559. doi:10.1016/S0022-3999(96)00228-0

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3), 185-216.

Brüne; M. (2005). "Theory of Mind" in Schizophrenia: A Review of the Literature, *Schizophrenia Bulletin*, Volume 31, Issue 1, 21–42, https://doi.org/10.1093/schbul/sbi002

Brunet , E., Sarfati, Y. & Hardy-Baylé, M.C. (2003) Reasoning about physical causality and other's intentions in schizophrenia, *Cognitive Neuropsychiatry*, 8:2, 129-139, DOI: 10.1080/13546800244000256

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M. & Zalla, T. (2013). The role of Causal and intentional judgments in moral reasoning in individuals with high functioning autism. *Journal of Autism and Developmental Disorders*. 43, 458–470. doi: 10.1007/s10803-012-1588-7

Byrne, R.M.J. (2012). Reasoning about intentions. Talk atthe International Conference on Thinking, London, UK.

Byrne, R.M.J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, 6(10), 426-431.

Byrne, R.M.J. (2016). Counterfactual Thought: From Conditional Reasoning to Moral Judgment. *Annual Review of Psychology*, 67

Capps, L., Kasari, C., Yirmiya, N., Sigman, M. (1993). Parental perception of emotional expressiveness in children with autism. *Journal of Consulting and Clinical Psychology*, 61, 475-484. doi: http://dx.doi.org/10.1037/0022-006X.61.3.475

Cardinale E. M., Finger E. C., Schechter J. C., Jurkowitz I. T. N., Blair R. J. R., Marsh A. A. (2014). "Moral status of an action influences its perceived intentional status in adolescents with psychopathic traits," *in Oxford Studies in Experimental Philosophy*, 1 eds Knobe J., Lombrozo T., NIchols S., editors. (Oxford: Oxford University Press), 131–151.

Catellani, P., & Bertolotti, M. (2014). The effects of counterfactual defenses on social judgments. European *Journal of Social Psychology*, 44(1), 82–92.

Catellani, P., & Milesi, P. (2001). Counterfactuals and roles: mock victims' and perpetrators' accounts of judicial cases. *European Journal of Social Psychology*, 31, 247-264. doi: 10.1002/ejsp.39

Catellani, P., & Milesi, P. (2005). When the social context frames the case: Counterfactuals in the courtroom. In D. Mandel, D. Hilton, & P. Catellani (Eds.), The psychology of counterfactual thinking (pp. 183–198). London: Routledge.

Catellani, P., Alberici, A. I., & Milesi, P. (2004). Counterfactual thinking and stereotypes: The nonconformity effect. *European Journal of Social Psychology*, 34, 421–436.

Cecchetto, C., Korb, S., Rumiati, RI. & Aiello, M. (2017) Emotional reactions in moral decision-making are influenced by empathy and alexithymia. *Social Neuroscience*. 1-15, https://doi.org/10.1080/17470919.2017.1288656

Ciaramelli, E., Braghittoni, D., di Pellegrino, G., (2012). It is the outcome that counts! Damage to the ventromedial prefrontal cortex disrupts the integration of outcome and belief information for moral judgment. *Journal of the international neuropsychological society*, 18, 962 – 971

Cokely, E.T., & Feltz, A. (2009). Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality*, 43, 18–24. doi: 10.1016/j.jrp.2008.10.007

Constantino, J. N. (2011). The quantitative nature of autistic social impairment. *Pediatric Research*, 69, 55–62. DOI:10.1203/PDR.0b013e318212ec6e

Corcoran, R., Cahill, C., & Frith, C. D. (1997). The appreciation of visual jokes in people with schizophrenia: A study of ‘‘mentalizing’’ ability. *Schizophrenia Research*, 24, 319-327.

Cova, F., Lantian, A. & Boudesseul, J. (2016). “Can the Knobe Effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality.” *Personality and Social Psychology Bulletin*. DOI: 10.1177/0146167216656356

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 353–80

Davis, C. G., Lehrnan, D. R., Wortman, C. B., & Silver, R. C. (1995). The undoing of traumatic life events. *Personality & Social Psychology Bulletin*, 21, 109-124.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10, 85.

Decety, J., & Moriguchi, Y. (2007). The empathic brain and its dysfunction in psychiatric populations: implications for intervention across different clinical conditions. *BioPsychoSocial Medicine (BPSM),* 1, 22–65. doi: 10.1186/1751-0759-1-22.

Díaz, R., Viciana, H., & Gomila, A., (2017). Cold Side-Effect Effect: Affect Does Not Mediate the Influence of Moral Considerations in Intentionality Judgments. *Frontiers in Psychology*; 8: 295. doi: 10.3389/fpsyg.2017.00295

Digman, J. M. (1990). Personality structure: emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417–440.

Dodell-Feder, D., Lincoln, SH., Coulson, J.P., & Hooker, C.I. (2013). Using Fiction to Assess Mental State Understanding: A New Task for Assessing Theory of Mind in Adults. *PLoS ONE*, 8(11):e81279. doi: https://doi.org/10.1371/journal.pone.0081279

Duff, A. R. (2015). Intention, intentional action and the law. *Methode-Analytic Perspectives*, 4(6), 136-146. doi:10.13135/2281-0498%2F128

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J. Brand, M., Kessler, J., Woike, J.K., Wolf, O.T. & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, 36, 623–636. doi:10.1007/s10803-006-0107-0

Dziobek, I., Rogers, K., Fleck, S. , Bahnemann, M., Heekeren, H.R., Wolf, O.T. & Covit, A. (2008). Dissociation of Cognitive and Emotional Empathy in

Adults with Asperger Syndrome Using the Multifaceted Empathy Test (MET) *Journal of Autism and Developmental Disorders*, 38, 464. doi:10.1007/s10803-007-0486-x

Eastabrook, J. M., Lanteigne, D. M. & Hollenstein, T. (2013). Decoupling between physiological, self-reported, and expressed emotional responses in alexithymia. *Personality and Individual Differ*ences 55**,** 978–982

Eichmann, M., Kugel, H., Suslow, T., (2008). Difficulty identifying feelings and automatic activation in the fusiform gyrus in response to facial emotion. *Perceptual and Motor Skills* 107*,* 915–922.

Epstude, K. & Roese, N.J. (2008). The function al theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2), 168–192.

Epstude K. & Roese N. J. (2011). When goal pursuit fails: The functions of counterfactual thought in intention formation. *Social Psychology*;42:19–27.

Fan, Y.T., Chen, C.Y., Chen, C.H., Decety, J. & Cheng, Y.W. (2014). Empathic arousal and social understanding in individuals with autism: evidence from fMRI and ERP measurements. *Social Cognitive and Affective Neuroscience*, 9 1203–1213. doi: 10.1093/scan/nst1017

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191. doi: doi:10.3758/BF03193146

Feldman Hall, O., Dalgleish, T and Mobbs, D. (2012).Alexithymia decreases altruism in real social decisions. *Cortex* 49, 899–904.doi: 10.1016/j.cortex.2012.10.015

Feltz, A. (2007). The Knobe Effect: A Brief Overview. *The Journal of Mind and Behavior*, 28, 265-277.

Feltz, A & Cokely, E.T., &. (2011). Individual differences in theory-of-mind judgments: Order effects and side effects. *Philosophical Psychology* 24(3):343-355 DOI10.1080/09515089.2011.556611

Feltz, A., Cokely, E. T., & Nadelhoffer, T. (2009). Natural compatibilism versus natural incompatibilism: Back to the drawing board. *Mind and Language*, 24, 1–23.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, MA: Addison–Wesley.

Fletcher, P. C. & Frith, C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*. 10, 48–58 (2009)

Forguson, L. (1989). Common sense. London: Routledge.

Forth, A.E., Kosson, D.S., and Hare, R.D. (2003). The Psychopathy Checklist: Youth Version. Toronto: Multi-Health Systems.

Franz, M., Schaefer, R. & Schneider, C. (2003). Psychophysiological Response Patterns of High and Low Alexithymics Under Mental and Emotional Load Conditions. *Journal of Psychophysiology.* 17**,** 203–213.

Fu, G., Xiao, W., Killen, M., & Lee K. (2014). Moral Judgment and Its Relation to Second-Order Theory of Mind. *Developmental Psychology*, 50, 2085–2092. doi: 10.1037/a0037077.

Funder, D. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science*, 2, 31–39.

Funder, D. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group Processes & Intergroup Relations*, 8, 109–124. doi: 10.1177/1368430205051060

Gambetti, E., Nori, R., Marinello, F., Zucchelli, M M., & Giusberti, F. (2017). Decisions about a crime: downward and upward counterfactuals; *Journal of Cognitive Psychology*, 1-12. doi:10.1080/20445911.2016.1278378

Georges L. C., Wiener R. L., Keller S. R. (2013). The angry juror: sentencing decisions in first degree murder. *Applied Cognitive Psychology*. 27 156–166. 10.1002/acp.2880

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

Gillespie, SM, Mitchell, I. J. and Abu-Akel, A.M. (2017). Autistic traits and positive psychotic experiences modulate the association of psychopathic tendencies with theory of mind in opposite directions. *Scientific Reports*, 7: 6485. doi: 10.1038/s41598-017-06995-2

Goerlich-Dobre, K.S., Lamm, C., Pripfl, J., Habel, U., Votinov, M. (2015).The left amygdala: A shared substrate of alexithymia and empathy. *Neuroimage*, 122:20-32. doi: 10.1016/j.neuroimage.2015.08.014.

Goerlich–Dobre K. S., Witteman J., Schiller N. O., van Heuven V. J. P., Aleman A., Martens S. (2013 Epub). Blunted feelings: alexithymia is associated with a diminished neural response to speech prosody. *Social Cognitive and Affective Neuroscience* , 9(8), 1108–17.

Gökçen, E., Petrides, K.V., Hudry, K., Frederickson, N., & Smillie, L.D. (2014).

Sub-threshold autism traits: The role of trait emotional intelligence and cognitive flexibility. *British Journal of Psychology*, 187–199. doi: 10.1111/bjop.12033.

Gökçen, E., Frederickson, N., & Petrides, K.V. (2016). Theory of mind and executive control deficits in typically developing adults and adolescents with high levels of autism traits. *Journal of Autism and Developmental Disorders*, 46, 2072–2087. doi: 10.1007/s10803-016-2735-3

Goldberg, J., Lerner, J., and Tetlock, P. (1999). Rage and reason: the psychology of the intuitive prosecutor. *European Journal of Social Psychology*. 29, 781–795.

Grabe, H. J., Wittfeld, K., Hegenscheid, K., Hosten, N., Lotze, M., Janowitz, D., Völzke, H., John, U., Barnow, S. and Freyberger, H. J. (2014), Alexithymia and brain gray matter volumes in a general population sample. *Human Brain Mapping*, 35: 5932–5945. doi:10.1002/hbm.22595

Grandi, S., Sirri, L., Wise, T.N., Tossani, E. and Fava, G.A. (2011). Kellner's Emotional Inhibition Scale: a clinimetric approach to alexithymia research. *Psychotherapy and Psychosomatics*. 80:335-344

Green S. R., Kragel P. A., Fecteau M. E. & LaBar K. S. (2014) Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, 91, 186–193

Grynberg, D., Chang, B., Corneille, O., Maurage, P., Vermeulen, N., Berthoz, S., Luminet, O., (2012). Alexithymia and the processing of emotional facial expressions (EFEs): systematic review, unanswered questions and further perspectives. *PLoS ONE* 7, e42429.

Gvozdic, K., Moutier, S., Dupoux, E. & Buon, M. (2016). Priming Children's Use of

Intentions in Moral Judgement With Metacognitive Training. *Frontiers in Psychology*, 7190. doi:10.3389/fpsyg.2016.00190

Guglielmo, S., Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 36, 1635-1647. doi:10.1177/0146167210386733.

Habib, M., Cassotti, M., Borst, G., Simon, G., Pineau, A., Houdé, O., & Moutier, S. (2012). Counterfactually mediated emotions: a developmental study of regret and relief in a probabilistic gambling task. *Journal of Experimental Child Psychology*, 112, 265-274.

Hadjikhani, N., Zürcher, N R., Rogier, O., Hippolyte, L., Lemonnier, E., Ruest, T., Ward, N., Lassalle, A., Gillberg, N., Billstedt, E., Helles, A., Gillberg, C., Solomon, P. & Prkachin, K. M. (2014). Emotional contagion for pain is intact in autism spectrum disorders. *Translational Psychiatry*, 4 (1): e343. doi:10.1038/tp

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998–1002.

Hayashi, H. (2007). Young children’s understanding of second-order mental states. *Psychologia*, 50,15–25. doi: 10.1111/j.1468-5884.2007.00352.x.2013.113.

Heckhausen, H. (1991). Motivation and action. Berlin: Springer Verlag.

Heider, F. (1958). The psychology of interpersonal relations. New York:Wiley.

Hill, E.L., Bird, C.A. (2006). Executive processes in Asperger syndrome: Patterns of performance in a multiple case series. *Neuropsychologia*, 44:2822–12835

Hindriks, F., Douven, I., Singmann, H. (2016). A new angle on the Knobe Effect: Intentionality correlates with blame, not with praise. *Mind & Language*, 31, 204–220.

Hodsoll, S., Viding, E., Lavie, N. (2011). Attentional capture by irrelevant emotional distractor faces. *Emotion,* 11, 346–353.

Hoekstra, R. A., Bartels, M., Cath, D. C. & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders,* 38, 1555–1566. doi:10.1007/ s10803-008-0538-x

Hughes, C., Jaffee, S.R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T.E. (2005). Origins of Individual Differences in Theory of Mind: From Nature to Nurture? *Child Development,* 76, 356-370. doi: 10.1111/j.1467-8624.2005.00850_a.x

Hughes, J.S., Trafimow, D. (2012). Inferences about character and motive influence intentionality attributions about side effects. *British Journal of Social Psychology* 51(4):661-73. doi: 10.1111/j.2044-8309.2011.02031.x.

Hughes, J.S., Trafimow, D. (2014). Mind attributions about moral actors: intentionality is greater given coherent cues. *British Journal of Social Psychology*. 54(2):220-35. doi: 10.1111/bjso.12077.

Ihme K., Sacher J., Lichev V., Rosenberg N., Kugel H., Rufer M., et al. (2014). Alexithymic features and the labeling of brief emotional facial expressions - an fMRI study. *Neuropsychologia* 64 289–299. 10.1016/j.neuropsychologia.2014.09.044

Jacobs, R.H., Renken, R., Aleman, A., Cornelissen, F.W. (2012). The amygdala, top-down effects, and selective attention to features. *Neuroscience and Biobehavioral Reviews* 36, 2069–2084

Jones, A. P., Happé, F.G., Gilbert, F., Burnett, S. & Viding, E. (2010). Feeling, caring, knowing: different types o f empathy deficit in boys with psychopathic Tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 51, 1188–1197. doi:10.1111/j.1469-7610.2010.02280.x

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 371–388). Hillsdale, NJ: Erlbaum

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-1 53.

Kano, M., Fukudo, S., 2013. The alexithymic brain: the neural pathways linking alexithymia to physical disorders. *Biopsychosocial medicine* 7, 1.

Kasimatis, M., & Wells, G. (1995). Individual differences in counterfactual thinking. In N. Roese & J. Olsen (Eds.), The social psychology of counterfactual thinking (pp. 81–101). Mahwah, NJ: Lawrence Erlbaum Associates.

Knobe, J. (2003a). Intentional Action and Side effects in Ordinary language. *Analysis*, 63, 190-193. doi: 10.1111/1467-8284.00419

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16, 309-323.

Knobe, J. (2004a). Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology*, 24, 270-279.

Knobe, J. (2004b). Intention, intentional action, and moral considerations. *Analysis*, 64, 181-187.

Knobe, J. (2006). The concept of intentional action: A case study in uses of folk psychology. *Philosophical Studies*, 130, 203-231.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences,* 33, 315-29. doi: 10.1017/S0140525X10000907.

Knobe, J., and Burra, A. (2006). The folk concept of intention and intentional action: A cross- cultural study. *Journal of Culture and Cognition*, 6, 113-132.

Knobe, J., and Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology,* 24, 252-258.

Kobick, J. (2010). Discriminatory intent reconsidered: Folk concepts of intentionality and equal protection jurisprudence. Harvard Civil Rights-Civil Liberties Law Review, 45, 517-562

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences* (*PNAS),* 110, 5648–5653. doi: 10.1073/pnas.1207992110.

Lagnado, D. A. & Channon, S. (2008). Judgments of Cause and Blame: The influence of Intentionality and Foreseeability. *Cognition*, 108, 754-770.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs, *Frontiers in Psychology,* 4: 863. doi: 10.3389/fpsyg.2013.00863

Larsen, J., Brand, N., Bermond, B., Hijman, R. (2003). Cognitive and emotional characteristics of alexithymia: a review of neurobiological studies. *Journal of Psychosomatic Research* 54, 533–541.

Leekam, S. R. (2016). Social cognitive impairment and autism: What are we trying to explain? *Philosophical Transactions of the Royal Society of London Series B Biological Science*s, 371 (1686), 20150082. DOI: 10.1098/rstb.2015.0082

Lerner J. S., Goldberg J. H., Tetlock P. E. (1998). Sober second thought: the effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin*. 24 563–574. 10.1177/0146167298246001

Leslie, A., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect. Theory of Mind and Moral Judgment. *Psychological Science*, 17, 5. doi: 10.1111/j.1467-9280.2006.01722.x

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529–566.

Linden, W., Lenz, J.W. and Stossel, C. (1996). Alexithymia, defensiveness and cardiovascular reactivity to stress. *Journal of Psychosomatic Research*. 41(6):575-83.

Lockwood, P.L., Bird, G., Bridge, M. & Viding, E. (2013). Dissecting empathy: high levels of psychopathic and autistic traits are characterized by difficulties in different social information processing domains. *Frontiers in Human Neuroscience,* 7,760. DOI:10.3389/fnhum.2013.00760

Loureiro, C. P., & de Hollanda Souza, D. (2013). The Relationship between Theory of Mind and Moral Development in Preschool Children. *Paidéia (Ribeirão Preto)*, 23, 93-101. doi: http://dx.doi.org/10.1590/1982-43272354201311

Luminet, O., Vermeulen, N., Demaret, C., Taylor, G.J., Bagby, R.M. (2006). Alexithymia and levels of processing: evidence for an overal deficit in remembering emotion words. *Journal of Research in Personality* 40, 713–733.

Machery E. (2008). The folk concept of intentional action: philosophical and experimental issues. *Mind and Language* 23 165–189. 10.1111/j.1468-0017.2007.00336.x

Macrae, C. N., Milne, A. B., & Griffiths, R. J. (1993). Counterfactual thinking and the perception of criminal behavior. *British Journal of Psychology*, 84, 221-226.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3, 23-48.

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture,* 6, 87–112.

Malle, B., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology,* 33, 101–121.

Malle, B., & Nelson, S. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law*, 21, 563-580. doi: 10.1002/bsl.554

Mandel, D. R. (2003). Effect of counterfactual and factual thinking on causal judgements. *Thinking & Reasoning*, 9, 245–265.

Mandel, D. R. & Dhami, M. K. (2005). “What I did” versus “what I might have done”: Effect of factual versus counterfactual thinking on blame, guilt, and shame in prisoners. *Journal of Experimental Social Psychology*, 41, 627–635.

Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 71, 450–463.

Mangelli, L., Semprini, F., Sirri, L., Fava, G.A. and Sonino, N. (2006). Use of the Diagnostic Criteria for Psychosomatic Research (DPCR) in a community sample. *Psychosomatics.* 47:143:146

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology,* 29, 87-109.

Markman, K. D., & McMullen, M. N. (2003). A reflection and evaluation model of comparative thinking. *Personality & Social Psychology Review*, 7, 244-267

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, 44, 421-428.

Marques, J. A., Quelhas, A. C., Juhos, C., Couto, M., Rasga, C. M., & Batalha, S. (2014). Counterfactual thinking: Study of the focus effect of scenarios and blame ascriptions to victim and perpetrator. *Análise Psicológic*a, 32, 355–385.

Mazza, M., Pino, M.C., Mariano, M., Tempesta, D., Ferrara, M., De Berardis, D., et al. (2014). Affective and cognitive empathy in adolescents with autism spectrum disorder. *Frontiers in Human Neuroscience*, 7, 8–791. doi: 10.3389/fnhum.2014.00791

McCrae, R., & Costa, P. (1990). Personality in Adulthood. New York: Guilford Press.

McGill, A. L., & Klein, J. G. (1995). Counterfactual and contrastive reasoning in explanations for performance: Implications for gender bias. In N. J. Roese & J. M. Olson (Eds.), What might have been: The social psychology of counterfactual thinking (pp. 333–351). Mahwah, NJ: Erlbaum.

McMullen, M. N., Markman, K. D., & Gavanski, I. (1995). Living in neither the best nor worst of all possible worlds: Antecedents and consequences of upward and downward counterfactual thinking. In N. J. Roese & J. M. Olson (Eds.), What might have been: The social psychology of counterfactual thinking (pp. 133- 167). Hillsdale, NJ: Erlbaum.

Mikhail, J. (2007).Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Science*. 11, 143–52

Minitab Statistical Software (2011). http://www.minitab.com/en-us/products/minitab/

Medvec, V., & Savitsky, K. (1997). When doing better means feeling worse: The effects of categorical cutoff points on counterfactual thinking and satisfaction. *Journal of Personality and Social Psychology*, 72, 1284–1296.

Mele, A., and Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31, 184–201. doi: 10.1111/j.1475-4975.2007.00147.x

Mellers, B.A., Schwartz, A., Ho K., & Ritov I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, 8, 423–429

Miller, G., & Johnson-Laird, P.N. (1976). *Language and perception*. Harvard: Harvard University Press.

Moore, J. W. & Pope. A. (2014). The intentionality bias and schizotypy. *The Quarterly Journal of Experimental Psychology*, 67, 11, 2218-2224. doi: https://doi.org/10.1080/17470218.2014.911332

Moriguchi, Y., Komaki, G., (2013). Neuroimaging studies of alexithymia: physical, affective, and social perspectives. *Biopsychosocial Medicine* 7, 8–8.

Moriguchi, Y., Ohnishi, T., Lane, R.D., Maeda, M., Mori, T., Nemoto, K., Matsuda, H., Komaki, G. (2006). Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. *Neuroimage* 32, 1472–1482.

Moss, D. (2017). Experimental Philosophy, Folk Metaethics and Qualitative Methods. *Teorema*, 36 (3), pp. 185-203. doi: 10.1080/09515089.2011.633751

Murphy, J. (2005). Individual differences in counterfactual thinking: The role of personality and health (Unpublished doctoral dissertation). Adelphi University, Long Island, New York.

Nadelhoffer, T. (2004a). Praise, side effects, and intentional action. *Journal of Theoretical and Philosophical Psychology*, 24, 196–213. doi: http://dx.doi.org/10.1037/h0091241

Nadelhoffer, T. (2004b). Blame, badness, and intentional: a reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, 24, 259-269.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9, 203-220.

Nanay, B. (2010). Morality of modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy*, 40, 28–40.

Nario-Redmond, M. R., & Branscombe, N. R. (1996). It could have been better or it might have been worse: Implications for blame assignment in rape cases. *Basic & Applied Social Psychology*, 18, 347-366.

Ndubuisi, B. & Byrne, R. (2013). *Intentionality and choice*. The Annual Meeting of the Cognitive Science Society. Web site:http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0364/index.html

Neumann, S. A., Sollers, J. J., Thayer, J. F. & Waldstein, S. R. Alexithymia predicts attenuated autonomic reactivity, but prolonged recovery to anger recall in young women. *International Journal of Psychophysiology*. 53, 183–195 (2004).

Newman, G.E., De Freitas, J., Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Journal of Cognitive Science*. 39(1):96-125. doi: 10.1111/cogs.12134.50.

Newton, T. L. & Contrada, R. J**. (**1994). Alexithymia and repression: contrasting emotion-focused coping styles. *Psychosomatic Medicine*.56(5):457-62.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, 22, 346-365. doi: 10.1111/j.1468-0017.2007.00312.x

Ngo, L., Kelly, M., Sinnott-Armstrong, W., Huettel, S. A., Coutlee, C. G., Carter, R. M., et al. (2015). Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports - Nature*. 5 1–11. 10.1038/srep17390

Nunez, M, Harris, PL. (1998). Psychological and deontic concepts: Separate domains or intimate connection? *Mind & Language*, 13:153–170

Nuñez, N., Schweitzer, K., Chai, C. and Myers, B. (2015). Negative Emotions Felt During Trial: The Effect of Fear, Anger, and. Sadness on Juror Decision Making, 29. *Applied cognitive psychology*. 200, 205

Ossorio, P. G., & Davis, K. E. (1968). The self, intentionality, and reactions to evaluations of the self. In C. Gordon & K. J. Gergen (Eds.), The self in social interaction. New York:Wiley

Ozonoff, S. (1997). Components of executive function in autism and other disorders. In J. Russell (Ed.), Autism as an executive disorder (pp. 179–211). Oxford, UK: Oxford University Press

Panaite, V. & Bylsma, L. M. (2012). In *Encyclopedia of Human Behavior* (*Second Edition*) (ed. Ramachandran, V. S.) 92–99 (Academic Press).

Parker, J.D., Taylor, G.J., Bagby, R.M., Acklin, M.W., (1993). Alexithymia in panic disorder and simple phobia: a comparative study. *American Journal of Psychiatry*150, 1105–1107.

Parker, P.D., Prkachin, K.M. and Prkachin, G.C. (2005) Processing of facial expressions of negative emotion in alexithymia: the influence of temporal constraint. *Journal of Personality*. 73(4):1087-107

Patil, I., Calò, M., Fornasier, F., Cushman, F., & Silani, G. (2017). The behavioral and neural basis of empathic blame. Scientific Reports, 7:5200. doi: 10.1038/s41598-017-05299-9

Patil, I., Calò, M., Fornasier, F., Young, L., & Silani, G. (2017). Neuroanatomical correlates of forgiving unintentional harms. *Scientific Reports*, 7:45967. doi:10.1038/srep45967

Patil, I., Melsbach, J., Hennig-Fast, K. & Silani, G. (2016). Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism. *Scientific Reports*, 6:23637. doi 10.1038/srep23637

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology*, 5:501. doi:10.3389/fpsyg.2014.00501

Piaget, J. (1932). The moral judgment of the child. New York: Harcourt, Brace.

Pellizzoni, S., Girotto, V., & Surian, L. (2010). Beliefs and moral valence affect intentionality attributions. *Review of Philosophy and Psychology*, 1, 2, 201-209

Pellizzoni, S., Siegal, M., & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology,* 45(1), 289-295.http://dx.doi.org/10.1037/a0014165

Pennington, B.F., Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry*.;37:51–127.

Peyroux, E., Strickland, B., Tapiero, I., & Franck, N. (2014). The intentionality bias in schizophrenia. *Psychiatry Research*, 219(3), 426–430. http://dx.doi.org/10.1016/j.psychres.2014.06.034.

Phelan M. T., Sarkissian H. (2008). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*. 138 291–298. 10.1007/s11098-006-9047-y

Phelan, M., and Sarkissian, H. (2009). Is the "trade-off hypothesis" worth trading for? *Mind and Language* 24, 164–180. doi: 10.1111/j.1468-0017.2008.01358.x

Phelps, E.A., LeDoux, J.E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*.20;48(2):175-87. Doi: 10.1016/j.neuron.2005.09.025

Pollatos, O., Schubö, A., Herbert, B. M., Matthias, E. & Schandry, R. (2008). Deficits in early emotional reactivity in alexithymia. *Psychophysiology* 45**,** 839–846

Pollatos, O., Werner, N.S., Duschek, S., Schandry, R., Matthias, E., Traut-Mattausch, E., Herbert, B.M. (2011). Differential effects of alexithymia subscales on autonomic reactivity and anxiety during social stress. *Journal of Psychosomatic Research*, 70, 525–533

Porcelli, P. and Rafanelli, C. (2010). Criteria for Psychosomatic Research (DPCR) in the medical setting. *Current Psychiatry Report*, 12:246-254.

Porcelli, P. and Sonino, N. (2007). Psychological Factors Affecting Medical Conditions: A New Classification for DSM-V. *Advances in Psychosomatic Medicine*. Basel, Karger. vol 28, pp I–X.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, 36, 1-16. doi: https://doi.org/10.1016/0010-0277(90)90051-K

Prkachin, G. C., Casey, C, Prkachin, KM. (2009). Alexithymia and perception of facial expressions of emotion. *Personality and Individual Differences* 46(4):412-417 DOI10.1016/j.paid.2008.11.010

Reker, M., Ohrmann, P., Rauch, A.V., Kugel, H., Bauer, J., Dannlowski, U., Arolt, V., Heindel, W., Suslow, T. (2010). Individual differences in alexithymia and brain response to masked emotion faces. *Cortex* 46, 658–667. doi: 10.1016/j.cortex.2009.05.008

Ritvo, R. A., Ritvo, E. R., Guthrie, D., Ritvo, M. J., Hufnagel, D. H., McMahon, W., Tonge, B., Mataix-Cols, D., Jassi, A., Attwood, T., Eloff, J. (2011). The Ritvo Autism Asperger diagnostic scale-revised (RAADS-R): A scale to assist the diagnosis of autism spectrum disorder in adults: An international validation study. *Journal of autism and developmental disorders*, 41, 1076-1089. doi: 10.1007/s10803-010-1133-5

Robbins, E., Shepard J., Rochat, P. (2017). Variations in judgments of intentional action and moral evaluation across eight cultures. *Cognition* 164 (2017) 22–30. doi: https://doi.org/10.1016/j.cognition.2017.02.012

Roedema, T.M., Simons, R.F. (1999). Emotion-processing deficit in alexithymia. *Psychophysiology*. 36(3):379-87.

Roese, N.J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, 66, 805–818.

Roese, N.J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121, 133–148.

Roese, N.J, & Olson, J.M. (1995). Outcome controllability and counterfactual thinking. *Personality and Social Psychology Bulletin*, 21, 620–628.

Roese, N.J., & Olson, J.M. (1997). Counterfactual thinking: The intersection of affect and function. In: Zanna, M.P. (Ed.) Advances in experimental social psychology. Academic Press; San Diego, CA: p.1-59.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O.T., Convit, A. (2007). Who Cares? Revisiting Empathy in Asperger Syndrome. *Journal of autism and developmental disorders,* 37,709–715 doi: 10.1007/s10803-006-0197-8

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition,* 108(3), 771–780. http://dx.doi.org/10.1016/j.cognition.2008.07.001.

Ryan, T. A. (1970). Intentional behavior. An approach to human motivation. New York: Ronald Press.

Sanna, L, Meier, S., & Turley-Ames, K. (1998). Mood, self-esteem, and counterfactuals: Externally attributed moods limit self-enhancement. *Social Cognition*, 16(2), 267–286.

Sarfati, Y., Hardy-Bayld, M.C., Besche, C., Widlöcher D. (1997). Attribution of intentions to others in people with schizophrenia: a non-verbal exploration with comic strips. *Schizophrenia Research.* 25, 199-209.

Scarpazza, C. (2015). Deficit in the Emotional Embodiment in Alexithymia, [Dissertation thesis], Alma Mater Studiorum Università di Bologna. Dottorato di ricerca in International phd program in cognitive neuroscience, 27 Ciclo. DOI 10.6092/unibo/amsdottorato/6840.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, 101, 632–652.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002) E-Prime Reference Guide. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review*, 84, 1–66. http://dx.doi.org/10.1037/0033-295X.84.1.1

Schwenck, C., Mergenthaler, J., Keller, K., Zech, J., Salehi, S., Taurines, R., Romanos, M., Schecklmann, M. Schneider, W., Warnke, A. & Freitag, C.M. (2012). Empathy in children with autism and conduct disorder: group-specific Profiles and developmental aspects. *Journal of Child Psychology and Psychiatry*, 53, 651–659. doi: 10.1111/j.1469-7610.2011.02499.x

Segura, S. Fernandez-Berrocal, P. & Byrne, R.M.J. (2002). Temporal and causal order effects in counterfactual thinking. *Quarterly Journal of Experimental Psychology*, 55, 1295-1305.

Senju, A., Southgate, V., White, S. & Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome. *Science*, 325, 883-885. doi: http://dx.doi.org/10.1126/science.1176170

Shamay-Tsoory, S.G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, 46(5), 668-77. doi: 10.1016/j.cortex.2009.04.008.

Shamay-Tsoory, S.G., Tomer, R., Berger, B.D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired ‘‘affective theory of mind’’ is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology*, 18, 55–67. doi: 10.1097/01.wnn.0000152228.90129.99

Shamay-Tsoory, S.G., Tomer, R., Yaniv, S., Aharon-Peretz, J. (2002). Empathy deficits in Asperger syndrome: a cognitive profile. *Neurocase*, 8(3):245-52. doi: 10.1093/neucas/8.3.245

Shaver, K. G. (1985). The attribution of blame. New York: Springer-Verlag.

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science*, 17, 97–108. doi:10.1037/h0080138

Singer, T., Critchley, H.D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, 13, 334-40. DOI: 10.1016/j.tics.2009.05.001.

Skorinko, J. L., Laurent, S., Bountress, K., Nyein, K. P., & Kuckuck, D. (2014). Effects of perspective taking on courtroom decisions. *Journal of Applied Social Psychology*, 44, 303–318. doi: 10.1111/jasp.12222

Skulmowski A., Bunge A., Cohen B. R., Kreilkamp B., & Troxler N. (2015). Investigating conceptions of intentional action by analyzing participant generated scenarios. *Frontiers in Psychology*, 5, 6, 1630, doi: 10.3389/fpsyg.2015.01630. eCollection 2015.

Slavny, R. J .M. and Moore, J.W. (2017). Individual Differences in the Intentionality Bias and its Association with Cognitive Empathy. *Personality and Individual Differences*, 122:104-108 doi: 10.1016/j.paid.2017.10.010

Smith, A. (2009). "The Empathy Imbalance Hypothesis of Autism: A Theoretical Approach to Cognitive and Emotional Empathy in Autistic Development".

The Psychological Record, 59, 3, doi: http://opensiuc.lib.siu.edu/tpr/vol59/iss3/9

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13, 290–312.

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition*, 95, 1-30. DOI: 10.1016/j.cognition.2003.12.004

South, M., Ozonoff, S. & Mcmahon, W. M. (2007). The relationship between executive functioning, central coherence, and repetitive behaviors in the high-functioning autism spectrum. *Autism*, 11, 437-451. doi: 10.1177/1362361307079606

Sripada, C. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*, 159-176. doi:10.1007/s11098-009-9423-5

Sripada, C. and Konrath, S. (2011), Telling More Than We Can Know About Intentional Action. *Mind & Language*, 26: 353–380. doi:10.1111/j.1468-0017.2011.01421.x

Starita, F., Làdavas, E., & di Pellegrino, G. (2016). Reduced anticipation of negative emotional events in alexithymia. *Scientific Reports*, 6, 27664.

Stone, V.E., Baron-Cohen, S. & Knight, R.T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, 10, 640-656. doi: 10.1162/089892998562942.

Stone, L.A. & Nielson, K. A. Intact Physiological Response to Arousal with Impaired Emotional Recognition in Alexithymia. (2001). *Psychotherapy Psychosomatic.* 70**,** 92–102

Summerville, A., Roese, N.J. (2008). Dare to compare: Fact-based versus simulation-based comparison in daily life. *Journal of Experimental Social Psychology*, 44, 664–671.

Sverdlik, S. (2004). Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology*, 34, 224-236.

Swart, M., Kortekaas, R., Aleman, A. (2009). Dealing with feelings: characterization of trait alexithymia on emotion regulation strategies and cognitive-emotional processing. *PLoS ONE* 4, e5751.

Tasso, A. (1999). Controfattuali e causalità in psicologia e filosofia. *Sistemi Intelligenti*, XI, 261–280.

Taylor, G.J. (2000). Recent developments in alexithymia theory and research. *The Canadian Journal of Psychiatry*,45(2):134-42.

Taylor, G. J., Bagby, R. M. & Parker, J. D. A. (1991).The Alexithymia Construct: A Potential Paradigm for Psychosomatic Medicine. *Psychosomatics* 32**,** 153–164

Taylor, G. J., Bagby, R. M. & Parker, J. D. (2003). The 20-Item Toronto Alexithymia Scale: IV. Reliability and factorial validity in different languages and cultures. *Journal of Psychosomatic Research* 55, 277–283

Taylor, G., Bagby, R.M. (2004). New trends in alexithymia research. *Psychotherapy and Psychosomatics* 73, 68–77.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675–735.

Turley, K. J., Sanna, L. J., & Reiter, R. L. (1995). Counterfactual thinking and perceptions of rape. *Basic &Applied Social Psychology*, 17, 285-303.

Turner, R. & Felisberti, F. M. (2017). Measuring mindreading : a review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology*, 8(47). doi: http://dx.doi.org/10.3389/fpsyg.2017.00047.

Tversky, A., & Kahneman, D. (1982). Judgments under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgments under uncertainty: Heuristics and biases (pp. 3–20). New York, NY: Cambridge University Press.

van der Velde, J. Servaas, M.N., Goerlich, K.S., Bruggeman, R., Horton, P., Costafreda, S.G., Aleman, A. (2013). Neural correlates of alexithymia: A meta-analysis of emotion processing studies. *Neuroscience & Biobehavioral Reviews*. 37,1774–1785

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The Reading the Mind in the Eyes test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry*, 18(4), 326–354. http://dx.doi.org/10.1080/13546805.2012.721728.

Wehmer, F., Brejnak, C., Lumley, M.A. and Stettner, L. (1995). Alexithymia and physiological reactivity to emotion-provoking visual scenes. *Journal of Nervous Mental Diseases*. 183, 351-357.

Weiner, B. (2001). Responsibility for social transgressions: An attributional analysis. In B. F. Malle, L. J.Moses, & D. A. Baldwin (Eds.), Intentions and intentionality: Foundations of social cognition (pp. 331–344). Cambridge, MA: MIT Press.

Wellman, H.W., & Phillips, A.T. (2001). Developing intentional understandings. In B.F. Malle, L.J. Moses, & D.A. Baldwin (Eds.), Intentions and intentionality: foundations of social cognition (pp.125-148). Cambridge, MA: MIT Press.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, 56(2), 161-169.

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality & Social Psychology*, 53, 421-430.

Wiener, R. L., Gaborit, M., Pritchard, C. C., McDonough, E. M., Staebler, C. R., Wiley, D. C., et al. (1994). Counterfactual thinking in mock juror assessments of negligence: A preliminary investigation. *Behavioral Sciences & the Law*, 12, 89-102.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 13, 103–128. doi: https://doi.org/10.1016/0010-0277(83)90004-5

Wingbermühle, E., Theunissen, H., Verhoeven, W.M.A., Kessels, R.P.C., Egger, J.I.M. (2012). The neurocognition of alexithymia: evidence from neuropsychological and neuroimaging studies. *Acta Neuropsychiatrica* 24, 67–80.

Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *PNAS*. 107 (15) 6753-6758; doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Adolphs, R., Tranel, D., Hauser, M. (2006). Does emotion mediate the relationship between an action’s moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*. 6 291–304. 10.1163/156853706776931312

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences (PNAS),* 104, 8235 – 8240. doi: 10.1073/pnas.0701408104

Young, L., & Phillips, J. (2011). The paradox of moral focus. Cognition, 119, 166-178.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology,* 24, 358–365. doi: http://dx.doi.org/10.1037/0012-1649.24.3.358

Zalla, T. & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophical Psychology*, 4, 681–698. doi: 10.1007/s13164-011-0048-1

Zalla, T. & Sperduti, M. (2013). The amygdala and the relevance detection theory of autism: an evolutionary perspective. *Frontiers in Human Neuroscience,* 7, 894. doi: 10.3389/fnhum.2013.00894

Zalla, T., Stopin, A., Ahade, S., Sav, A.M. & Leboyer, M. (2009). Faux pas detection and intentional action in Asperger syndrome. A replication on a French sample. *Journal of autism and developmental disorders*, 39, 373–382.

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of Behavioural Decision Making*, 12, 93-106.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, 67, 2478–2492. doi: 10.2307/113163

Appendix A

Scenarios used in Study 3

The CEO started a plan to increase profits.

She did not care at all about the effect the plan would have on the environment.

She knew her plan would harm the environment.

Did the CEO intentionally harm the environment?

The chairman started a plan to increase revenue.

He did not care at all about the effect the plan would have on the environment.

He knew his plan would help the environment.

Did the chairman intentionally help the environment?

Jenny spread weed killer to protect her crops.

Jenny did not care at all about the effect this would have on Susie-Ann’s crops.

Jenny knew her pesticide would harm Susie-Ann's crops.

Did Jenny intentionally harm her neighbor's crops?

Stanley spread anti-fungals to protect his crops.

Stanley did not care at all about the effect this would have on Billy-Bob’s crops.

Stanley knew his anti-fungals would protect Billy-Bob’s crops.

Did Stanley intentionally protect Billy-Bob’s crops?

The scientist released a drug to gain profit.

She did not care at all about the effect the drug would have rates of cancer.

She knew her drug would increase the rate of cancer.

Did the scientist intentionally increase the rate of cancer?

The scientist released a drug to make a deadline.

He did not care at all about the effect the drug would have on rates of heart attacks.

He knew his drug would decrease the rate of heart attacks.

Did the scientist intentionally decrease the rate of heart attacks?

Rebecca protests in support of political prisoners to get on TV.

Rebecca did not care at all about the effect her protests would have on the prisoners.

Rebecca knew her protests would cause the execution of the political prisoners.

Did Rebecca intentionally cause the execution of the political prisoners?

Sean protests in support of the death row inmate to get onto newspapers.

Sean did not care at all about the effect this would have on the death row inmate.

Sean knew his protests would cause the release of the death row inmate.

Did Sean intentionally cause the release of the death row inmate?

Curtis released the documents to gain publicity.

Curtis did not care at all about the effect this would have on his friend's reputation.

Curtis knew the documents would ruin his friend's reputation.

Did Curtis intentionally ruin his friend's reputation?

Lori released the photos to gain news coverage.

Lori did not care at all about the effect this would have on her boss’s reputation.

Lori knew the photos would redeem her boss’s reputation.

Did Lori intentionally redeem her boss’s reputation?

The doctor prescribed the Elixir drug to make the drug company happy.

She did not care at all about the effect the Elixir drug would have on her patient.

She knew the Elixir drug would cause fatal bleeding for her patient.

Did the doctor cause the patient to have fatal bleeding?

The doctor prescribed the Gastropurge drug to please the drug vendor.

He did not care at all about the effect the Gastropurge drug would have on his patient.

He knew the Gastropurge drug would finally cure his patient.

Did the doctor intentionally cure his patient?

The mayor diverted water to Oldtown to gain votes.

He did not care at all about the effect this would have on Newtown.

He knew diverting the water would deprive Newtown of water.

Did the mayor intentionally deprive Newtown of water?

The councilwoman diverted water to her town to win an election.

She did not care at all about the effect this would have on food production.

She knew diverting the water for his town would increase food production.

Did the councilwoman intentionally increase food production?

The university president enacted a plan to increase business school funding.

He did not care at all about the effect the plan would have on the medical school.

He knew his plan would cut funding for the medical school.

Did the university president intentionally cut funding for the medical school?

The athletic director enacted a plan to increase basketball funding.

She did not care at all about the effect the plan would have on the soccer team.

She knew her plan would boost funding for the soccer team.

Did the athletic director intentionally boost funding for the soccer team?

Kate placed her uncle in a nursing home to avoid being his caretaker.

Kate did not care at all about the effect the placement would have on her uncle.

Kate knew placing him in a nursing home would make him extremely unhappy.

Did Kate intentionally make her uncle unhappy?

Jared placed his aunt in a nursing home to avoid being her caretaker.

Jared did not care at all about the effect the placement would have on his aunt.

Jared knew placing his aunt in a nursing home would make her extremely happy.

Did Jared intentionally make his aunt happy?

Jacob threw a party to be more popular.

Jacob did not care at all about the effect this would have on his roommate, Curtis.

Jacob knew his party would make Curtis fail the morning's exam.

Did Jacob intentionally make Curtis fail the morning's exam?

Rachel threw a party to have fun.

Rachel did not care at all about the effect this would have on her roommate, Jackie.

Rachel knew her party would help Jackie make new friends.

Did Rachel intentionally help Jackie make new friends?

Tina built a road to increase traffic to her country store.

Tina did not care at all about the effect this would have on a nearby 1000-year-old tree.

Tina knew her new road would cause the 1000-year-old tree to die.

Did Tina intentionally cause the 1000-year-old tree to die?

Ronald bought a trolley to increase visitors to his theme park.

Ronald did not care at all about the effect this would have on a nearby monument.

Ronald knew his new road would make the monument famous.

Did Ronald intentionally make the monument famous?

Norman protested in front of the church to express his views.

Norman did not care at all about the effect the protests would have on the church.

Norman knew his messages would offend the church.

Did Norman intentionally offend the church?

Angela protested in front of the church to express her views.

Angela did not care at all about the effect the protests would have on the church.

Angela knew her messages would earn the church good publicity.

Did Angela intentionally earn the church good publicity?

Other questionnaire used in Study 3

The Autism Quotient. The Autism Quotient (Baron-Cohen, Wheelwright, Skinner, et al., 2001) is a 50-item questionnaire measuring ASD traits in adults of normal intellectual ability. It comprises five dimensions corresponding to the social-cognitive deficits associated with ASD: social skills, attention-switching, attention to detail, communication, and imagination. The AQ requires individuals to indicate whether they ‘strongly agree’, ‘slightly agree’, ‘slightly disagree’, or ‘strongly disagree’ with each item. Individuals score in the range of 0–50, with higher scores reflecting more severe symptomatology. The AQ has been shown to distinguish between groups of individuals with ASD and age-matched controls (Baron-Cohen, Wheelwright, Skinner, et al., 2001). A score of 32 was considered as indicating "clinically significant levels of autistic traits", according to what was suggested in previous studies (Baron-Cohen et al., 2001). Test–retest and inter-rater reliability of the AQ were good, and a cut-off score of 32 has acceptably high sensitivity (0.77) and specificity (0.74) (Austin 2005; Woodbury-Smith et al. 2005). (Table 10)

	Autistic traits of personality		**Control group**	
Task	M	SD	M	SD
Autism Quotient	37,50	3,17	18,70	3,02
Autism and Asperger Diagnostic Scale Revised	84,02	15,22	41,78	14,52
Negative Side effect	73,41	17,96	63,71	21,39
Positive side effect	45,08	19,08	36,36	17,51

Table 10. Mean and standard deviation of Autism Quotient (from Baron-Cohen, Wheelwright, Skinner, et al., 2001); Autism and Asperger Diagnostic Scale Revised (from Ritvo et al., 2011), and of intentionality judgments provided to the negative and positive side effects of like–Knobe scenarios (from Ngo et al., 2015).

Appendix B

Scenarios used in study 4

Scenarios characterized by high salience content

Bill released a gadget to make a deadline.

Bill did not care at all about the effect the gadget would have on babies.

Bill knew his gadget would kill babies.

Did Bill intentionally cause the death of babies?

Robyn released an invention to make a deadline.

Robyn did not care at all about the effect the invention would have on toddlers.

Robyn knew her invention would help toddlers.

Did Robyn intentionally help toddlers?

The airplane bomber bombed a factory to reduce enemy's steel production.

He did not care at all about the effect the bombing would have on innocent civilians.

He knew his bombing would kill innocent civilians.

Did the airplane bomber intentionally kill innocent civilians?

The bomber pilot bombed a facility to reduce the enemy's iron production.

She did not care at all about the effect the bombing would have on the townsfolk.

She knew her bombing would liberate the townsfolk.

Did the bomber pilot intentionally liberate the townsfolk?

The scientist released a drug to gain profit.

She did not care at all about the effect the drug would have rates of cancer.

She knew her drug would increase the rate of cancer.

Did the scientist intentionally increase the rate of cancer?

The scientist released a drug to make a deadline.

He did not care at all about the effect the drug would have on rates of heart attacks.

He knew his drug would decrease the rate of heart attacks.

Did the scientist intentionally decrease the rate of heart attacks?

Rebecca protests in support of political prisoners to get on TV.

Rebecca did not care at all about the effect her protests would have on the prisoners.

Rebecca knew her protests would cause the execution of the political prisoners.

Did Rebecca intentionally cause the execution of the political prisoners?

Sean protests in support of the death row inmate to get onto newspapers.

Sean did not care at all about the effect this would have on the death row inmate.

Sean knew his protests would cause the release of the death row inmate.

Did Sean intentionally cause the release of the death row inmate?

The doctor prescribed the Elixir drug to make the drug company happy.

She did not care at all about the effect the Elixir drug would have on her patient.

She knew the Elixir drug would cause fatal bleeding for her patient.

Did the doctor cause the patient to have fatal bleeding?

The doctor prescribed the Gastropurge drug to please the drug vendor.

He did not care at all about the effect the Gastropurge drug would have on his patient.

He knew the Gastropurge drug would finally cure his patient.

Did the doctor intentionally cure his patient?

Scenarios characterized by standard salience content

The CEO started a plan to increase profits.

She did not care at all about the effect the plan would have on the environment.

She knew her plan would harm the environment.

Did the CEO intentionally harm the environment?

The chairman started a plan to increase revenue.

He did not care at all about the effect the plan would have on the environment.

He knew his plan would help the environment.

Did the chairman intentionally help the environment?

Jenny spread weed killer to protect her crops.

Jenny did not care at all about the effect this would have on Susie-Ann’s crops.

Jenny knew her pesticide would harm Susie-Ann's crops.

Did Jenny intentionally harm her neighbor's crops?

Stanley spread anti-fungals to protect his crops.

Stanley did not care at all about the effect this would have on Billy-Bob’s crops.

Stanley knew his anti-fungals would protect Billy-Bob’s crops.

Did Stanley intentionally protect Billy-Bob’s crops?

Tina built a road to increase traffic to her country store.

Tina did not care at all about the effect this would have on a nearby 1000-year-old tree.

Tina knew her new road would cause the 1000-year-old tree to die.

Did Tina intentionally cause the 1000-year-old tree to die?

Ronald bought a trolley to increase visitors to his theme park.

Ronald did not care at all about the effect this would have on a nearby monument.

Ronald knew his new road would make the monument famous.

Did Ronald intentionally make the monument famous?

Curtis released the documents to gain publicity.

Curtis did not care at all about the effect this would have on his friend's reputation.

Curtis knew the documents would ruin his friend's reputation.

Did Curtis intentionally ruin his friend's reputation

Lori released the photos to gain news coverage.

Lori did not care at all about the effect this would have on her boss's reputation.

Lori knew the photos would redeem her boss's reputation.

Did Lori intentionally redeem her boss's reputation?

Norman protested in front of the church to express his views.

Norman did not care at all about the effect the protests would have on the church.

Norman knew his messages would offend the church.

Did Norman intentionally offend the church?

Angela protested in front of the church to express her views.

Angela did not care at all about the effect the protests would have on the church.

Angela knew her messages would earn the church good publicity.

Did Angela intentionally earn the church good publicity?

Acknowledgements

At the end of this experience, I would like to give thanks to all those who helped me along the way.

First of all, I would like to thank my tutor, Prof. Giusberti for giving me the opportunity to explore this fascinating area at the crossroads of Psychology, Philosophy, Neuroscience, Law and many more. And thanks to the CNC’s team with whom I was able to carry out the latest study of this research.

A big thank you to my colleagues Raffaella and Elisa for being always available in these years, and above all, for creating a relationship of trust beyond the working environment.

I would also like to thank those who facilitated and enriched my period abroad, an experience the Ph.D. allowed me to undertake: I sincerely thank Indrajeet for the period spent working with him, which was extremely stimulating; I thank Professor Fiery Cushman for welcoming me to Harvard University with the utmost willingness and kindness; I thank Indrajeet, the Professor and Giuseppe for the collaborations we started, which gave a further important contribution to my professional experience.

In general I would like to thank all those outside the University, especially Susanna, who made me feel like I was at home despite the distance.

At home, a special thanks goes to my grandmother Laura, who with great commitment and dedication helped me look for the first participants and begin this research.

Finally, a huge thanks to the rest of my family and to Marco for always supporting me in all my experiences.