Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-025-00792-x

Assessing the Knobe Effect in Autistic and Non-Autistic Individuals

José V. Hernández-Conde[1] · **Agustín Vicente**[2,3]

Received: 6 October 2024 / Accepted: 30 October 2025

Abstract
The present study embarks on a comprehensive investigation of whether the influence of the moral valence, as highlighted by Knobe, remains a predominant factor in subjects' attributions of intentionality across diverse scenarios and populations. In addition to examining the harm dimension, our research explores the potential presence of this effect in other circumstances, assessing whether there is a comparable influence on attributions of intentionality for cases with side effects not describable as harmful. A comparative analysis between neurotypical and autistic populations is conducted –-in line with Zalla & Leboyer's work, but extending the study to other scenarios–-, revealing some disparities in how these two groups attribute intentionality and praise and blame. Final analyses were conducted on the gender variable, which revealed significant gender differences within both populations.

1 Introduction

1.1 The Knobe Effect: Moral Judgment and Intentionality Attribution

The Knobe effect (Knobe 2003a, 2003b) describes a robust phenomenon whereby people tend to attribute intentionality to actions depending on whether the outcome of such actions is perceived as positive or negative. In particular, people are more likely to attribute intentionality to actions with negative moral valence (such as harm-

✉ José V. Hernández-Conde
jhercon@uva.es

✉ Agustín Vicente
agustin.vicente@ehu.eus

[1] University of Valladolid, Valladolid, Spain

[2] University of the Basque Country, UPV/EHU, Vitoria-Gasteiz, Spain

[3] Ikerbasque: Basque Foundation for Science, Bilbao, Spain

Published online: 04 December 2025

ing someone) than to actions with positive moral valence (such as helping someone) (Knobe 2003a, 2003b). This suggests that moral valence plays a crucial role in people's judgments of intentionality, and that people's perceptions of the moral character of an action's outcome can influence their attribution of intentionality to the agent.

Typically, the Knobe effect is tested by presenting participants with two different scenarios in which agents make a decision based on their own interests. In one scenario, they make their decision knowing about some harmful side effects (e.g. polluting the environment); in the other, they make their decision knowing about some beneficial side effects (e.g., improving the environment). Although the scenario narratives explicitly state that agents do not care about the side effects of their decisions, it is a robust result that participants ascribe a much higher degree of intentionality in the case of harmful side effects than in the case of beneficial side effects. However, little attention has been paid to whether this type of effect also occurs in cases where there is no harm, but only a side effect that can be judged as bad (Knobe 2004), and where it has been done, it has typically been to examine cases where an action could have resulted in a harmful consequence, but that harm did not ultimately occur (Cushman 2008).

Interestingly, the experimental paradigm used to test the Knobe effect includes a question about blame and praise. Participants are asked whether the agents whose intentionality they are judging should be blamed or praised for what they have done. Some results show that blame and praise attributions do not correlate with intentionality attributions (Knobe 2003a, 2003b; Pettit and Knobe 2009). In particular, several studies (e.g., Knobe 2003a, 2004; Nadelhoffer 2004; Guglielmo and Malle 2010) have identified a significant correlation between judgments of intentionality and moral blame, whereas such a relationship appears considerably weaker or altogether absent in cases involving moral praise. These findings suggest that the link between moral judgment and intentionality is modulated by the valence of the outcome (negative vs. positive).The Knobe effect is puzzling on its own, but it is even more puzzling when praise/blame judgments are compared with attributions of intentionality.

Despite the many explanations put forward, no consensus has yet emerged on the cause of the Knobe effect[1]. The *moral-valence* interpretation originally proposed by Knobe (2003a, 2003b) –which holds that people's moral evaluation of the side effect (i.e., whether they deem it positive or negative) heavily shapes their judgments of intentionality– remains the leading explanation. Evidence for this claim comes not only from the persistence of the asymmetry when "intend to" is replaced by other verbs, such as "decide," "defend," or "oppose" (Pettit and Knobe 2009), but also from the robust replication of the phenomenon across a wide range of subsequent studies (Young et al. 2006; Ngo et al. 2015; Phillips et al. 2015; Cova et al. 2016; Stewart et al. 2022). Another early explanation was that the asymmetry was due to *different conceptions of intentional action* (Nichols and Ulatowski 2007; Cushman and Mele 2008; Laurent et al. 2021), one based on desires (to cause the side effect) –when subjects do not attribute intentionality to the agent– and the other based on beliefs (that the side effect would occur) –when subjects do attribute intentionality.

[1] For a detailed review of the most influential interpretations, see Cova 2015.

In addition, there are also *pragmatic explanations* that suggest that when subjects attribute intentionality to the agent, they are expressing their disapproval of the agent's decision through a conversational implicature (Adams and Steadman 2004a, 2004b); *normativist explanations*, which suggest that the greater intentionality attributed in negative cases is due to the fact that they are instances of norm violation –either because people form stronger beliefs about negative collateral consequences (Holton 2010; Uttich and Lombrozo 2010; Alfano et al. 2012), or because when agent's actions violate established norms they are often seen as acting more freely and deliberately, as the primarily responsible for the resulting consequences (Kirfel and Phillips 2023)–; and *probabilistic explanations*, which claim that the asymmetry arises from the different perceptions of the probabilities that the agent's action will produce the collateral effect for positive and negative cases (Dalbauer and Hergovich 2013; Nakamura 2018).

In recent years, Knobe (2022) has proposed what he calls the *possibility hypothesis*, according to which people's intentionality judgements are shaped by their moral considerations about the situation under consideration, but are also modulated by which possibilities they regard as relevant. Other factors that have been observed to influence judgments of intentionality are carefulness in acting and foreseeability (Margoni & Surian 2022) –the latter in relation to the possibility hypothesis– as well as personality traits and emotions (Young et al. 2006; Feltz and Cokely 2024). Emotion activation has been shown to increase intentionality judgments for negative side effects, especially when there is an immediate intuitive response (e.g., Ng o et al. 2015; Zucchelli et al. 2019; Zuchelli et al. 2025). Zucchelli and collaborators (2019) observed a reduction of the effect in individuals exhibiting alexithymia, a condition consisting in difficulties in identifying, describing, and processing one's own emotions.

The consistent and widespread occurrence of the Knobe effect across a variety of contexts provides a compelling rationale for examining whether analogous asymmetries in intentionality attribution may exist beyond the harm domain. According to some moral psychologists (Graham et al. 2009, 2018) harm has been turned into the most relevant moral dimension in current liberal societies. It is therefore pertinent to investigate whether outcome valence –whether positive or negative– elicits comparable asymmetries in intentionality attribution when the evaluative dimension at stake relates to the aesthetic or alethic domains, and whether such effects can occur independently of moral considerations. Associated with this investigation, two central research questions emerge: Does the Knobe effect extend to aesthetic and alethic evaluations, and if so, can these effects be empirically disentangled from moral influences? Is the magnitude of the asymmetry consistent across these three domains, or do domain-specific variations emerge that would point to distinctive evaluative processes?

On this basis, a central objective of the present work is to carry out a comparative analysis of intentionality attribution across the harm, aesthetic, and alethic domains, with the aim of clarifying the scope and boundaries of the side-effect effect and assessing whether the observed asymmetry in intentionality attribution is intrinsically rooted in moral cognition or, alternatively, reflects a broader evaluative bias that extends across multiple dimensions of human judgment.

1.2 Autism: Morals and Intentionality Attribution

Autism is characterized as a neurodevelopmental profile that departs from the neurotypical profile in several ways: autistic individuals tend to exhibit more sensitivity to the environment and to be more easily overwhelmed by environmental stimuli than neurotypicals; they also exhibit more difficulties in social interaction and communication, and show a stronger preference than neurotypicals for predictable environments and routines, and for in-depth exploration of their own interests. Autistic individuals are said to exhibit three basic cognitive differences from neurotypicals: stronger local processing (Happé & Frith 2006), more difficulties in the area of executive functioning (Hill 2004), and more difficulties in the area of theory of mind abilities (Baron-Cohen 2001).

Compared to other areas of research in autism, little work has been done in the area of moral development and moral psychology in autistic individuals. Concerning moral development, autistic children have been found to be less strict about the moral/conventional distinction than neurotypical children (Shulman et al. 2012; but see Blair 1996) and less elaborate in their judgments (Shulman *et al. ibid.*). They have also been found to have difficulty distinguishing bad outcomes of intentional vs. accidental actions from a moral perspective, such that they tend to judge more in terms of outcomes than in terms of intentions (Moran et al. 2011; Gleichgerrcht and Young 2013; Margoni and Surian 2016). This pattern of behavior is thought to be related to theory of mind difficulties (Garcia-Molina and Clemente-Estevan 2019).

Since theory of mind difficulties in autism are thought to persist into adulthood, it is assumed that autistic individuals will show a stronger tendency than neurotypicals to judge the rightness or wrongness of an action by its consequences, especially in cases where tracking the intentions of agents may be complicated (as in cases of deceit or intentional misleading: see Garcia-Molina and Clemente-Estevan 2019). So far, studies with autistic adults have provided some support for this idea. For example, Zalla and Leboyer (2011) tested 20 autistic adults for the Knobe effect and found that their blame/praise judgment was less related to their attribution of intentionality than in the neurotypical case, suggesting that the autistic participants did not evaluate agents' actions on the basis of their intentions. Zucchelli and collaborators (2018) investigated the interplay between autistic personality traits, cognitive and affective theory of mind capacities, and the attribution of intentionality. The findings revealed that elevated levels of autistic traits were significantly associated with diminished cognitive and affective theory of mind abilities, alongside an increased tendency to attribute intentionality to side effects of actions. Moreover, the study identified cognitive theory of mind difficulties as a mediating factor in the relationship between autistic traits and the heightened attribution of intentionality, again suggesting that individuals exhibiting high autistic traits are more inclined to evaluate intentionality based on the occurrence of side effects rather than the agent's underlying intentions. Finally, Machery and Zalla (2014) report results showing that autistic adults judge that actions that are merely instrumental in bringing about an end are unintentional.

On the other hand, autistic individuals may exhibit a stronger sense of justice or more consistent morals than neurotypicals (Dempsey et al. 2020; Hu et al. 2021).

Some studies suggest that autistic individuals are typically more "legalistic" than neurotypicals (Strang et al. 2017; Dempsey *et al. ibid.*). In turn, neurotypicals are more prone to accept exceptions to moral norms. First-person accounts from autistic adults suggest that many are irritated by what they perceive as the moral laxity of neurotypicals (Hu et al. 2021).

In this paper we contribute to the literature on intentionality attribution and on moral judgment in autism. We introduced 198 participants (99 autistic, 99 non-autistic) to three different Knobe cases, one about harm/help, one about false/true information, and a final one about good/bad aesthetic outcomes. Participants were asked whether the outcomes of the actions were intentional and whether the agents were to be blamed or praised for what they did. We used several scenarios because in a previous study, focused on interdomain differences, we had observed that neurotypical individuals did not react in the same way to the three scenarios, apparently making their responses dependent on the severity of the side-effect. Thus, we were interested not only in measuring the reactions of autistic individuals to the original Knobe effect, but also in comparing their responses with those of neurotypicals in cases where the side effects do not consist in harm, even if the decision taken by the agent is prima facie immoral.

Our hypotheses regarding the results of the comparison between autistic and non-autistic adult participants were:

(1) There will be differences in the moral evaluation of the agents in our scenarios. Based on the above studies (Dempsey et al. 2020; Hu et al. 2021), we expect to observe higher standards in the praise/blame question in the autistic group. The question asked in the Knobe scenarios, in a nutshell, is whether the agent was right to act on the basis of the shellfish motivations and regardless of the known side effects. We expect that the type of autistic population tested will be able to understand the narrative, and that they will be stricter in their judgments than neurotypicals.
(2) Furthermore, in line with previous findings (Zalla and Leboyer 2011; Zucchelli et al. 2018), we may observe a larger discrepancy between praise/blame judgments and intentionality attributions in the autistic group than in the neurotypical group, as well as a larger Knobe effect. As mentioned above, it is still controversial why the Knobe effect occurs. However, the Knobe effect shows that in some cases people do not make a clear distinction between goals and side effects. Given that the literature to date suggests a developmental delay in distinguishing between incidental and intentional outcomes, and some persistent differences in means-ends evaluations (Moran et al. 2011; Gleichgerrcht and Young 2013; Margoni and Surian 2016), we expect to observe that autistic individuals will be more prone to the Knobe effect.
(3) Supposing that (1) and (2) were the case, the picture that would emerge is that autistic individuals may base their intentionality attributions on outcomes, but not their moral judgments –contrary to some of the previously cited literature (Moran et al. 2011; Gleichgerrcht and Young 2013; Margoni and Surian 2016).

2 Method

2.1 Participants

In our experiment, we recruited 99 neurotypical participants on a voluntary basis, including both university students and professors (65% female, predominantly native Spanish speakers, average age = 37) from three public universities in Spain.

We also recruited 99 autistic participants on a voluntary basis through the Prolific platform (49% female, native English speakers, average age = 37). Each participant received 3£ as compensation for their participation in the study. These participants had earlier responded "Yes–as a child" or "Yes–as an adult" to the Prolific screener question, "Have you received a formal clinical diagnosis of autism spectrum disorder, made by a psychiatrist, psychologist, or other qualified medical specialist? This includes Asperger's syndrome, Autism Disorder, High Functioning Autism, or Pervasive Developmental Disorder".

2.2 Experiment Description

Experiment 1 reproduces Knobe's (2003a) experiment –for the *harm* domain–, which is also one of those examined by Zalla and Leboyer (2011) in their work about moral evaluation in "high-functioning" autism, and two other cases inspired by Knobe's original work –for the *aesthetic* and *alethic* domains. For these three domains two critical scenarios were used, namely, one with a positive side effect and one with a negative side effect. Participants were then asked to indicate, on a seven-point Likert scale (ranging from −3 to +3), the extent to which they agreed or disagreed with the statement "*agent* X acted well by *making decision* Y" for each scenario (where *agent* X and *decision* Y take specific and different values from one scenario to another), and whether they thought *agent* X intentionally contributed to the side effect presented.

Participants were exposed to one critical scenario from each of these three domains, along with six filler scenarios that acted as distractors[2]. These nine scenarios were presented in three consecutive blocks (i.e., harm, aesthetic, and alethic), each of which consisted of three scenarios: one critical scenario (of the harm, aesthetic, or alethic type, with either a positive or a negative side effect), and two distractor scenarios.

The order of presentation of the three blocks was randomized. The internal order of presentation of the three scenarios in each block and the assignment of the six distractor scenarios to the blocks was also randomized. All scenarios were presented in English.

The scenarios for the harm domain were the *help* and *harm* versions of Knobe's (2003a) original ENVIRONMENT vignette, which were also studied by Zalla and Leboyer (2011) in their article:

[2] The distractor scenarios were designed to resemble the critical scenarios in form, but without the collateral effects that are typically associated with cases involving the Knobe effect. See Appendix 1 for a detailed description of each of the six distractor scenarios.

ENVIRONMENT: 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/but* it will also *help/harm* the environment." The chairman of the board answered, "I don't care at all about *helping/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 helped/harmed.

And then the participants were asked:

> [1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:
> The chairman did the right thing by starting the new program.
> [2] Did the chairman intentionally*help/harm* the environment?

The other two scenarios, for the aesthetic and alethic domains, were closely modeled on Knobe's (2003a) scenarios, although they were adapted so that the dominant dimension in them did not result in a harm effect, but in an aesthetic or alethic/truth effect.

For the aesthetic domain, a STREET-FURNITURE vignette was designed, in which the agent (i.e., the mayor of the city) makes a decision that has either a positive or negative side effect on the overall aesthetics of the city:

STREET-FURNITURE: The urban planning councilor of a city went to the mayor and said, "We are thinking of replacing the street furniture with new ones. It will give the city a more modern look, but it will also *improve/worsen* the general aesthetics of the city." The mayor of the city answered, "I do not care at all about *improving/worsening* the aesthetics of the city. I just want it to look more modern. Let's install the new street furniture." The city council replaced the old street furniture with the new ones. Sure enough, the general aesthetics of the city *improved/worsened*.

And then the participants were asked:

> [1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:
> The mayor did the right thing by replacing the old street furniture with the new ones.
> [2] Did the mayor intentionally *improve/worsen* the aesthetics of the city?

Finally, in the alethic domain, participants were presented with an ECONOMIC-OUTLOOK vignette in which the agent (i.e., the candidate of a political party) makes a decision that has either a positive or negative side effect on the voter's perception of the economic outlook:

ECONOMIC-OUTLOOK: The campaign manager of a political party went to the candidate and said, "We are thinking of using a new argument in the campaign. It will help us respond to the opposition, but it will also make people get the *right/wrong* idea about the economic outlook." The candidate answered, "I do not care at all about voters getting the *right/wrong* idea about the economic outlook. I just want to

respond to the opposition. Let's use the new argument." The campaign staff used the new argument. Sure enough, the voters got the *right/wrong* idea about the economic outlook.

And then the participants were asked:

> [1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:
> The candidate did the right thing by using the new argument.
> [2] Did the candidate intentionally contribute to the voters getting the *right/wrong* idea about the economic outlook?

Additionally, we aimed to ensure that participants understood the new scenarios (i.e., the STREET-FURNITURE and ECONOMIC-OUTLOOK vignettes) in a manner analogous to Knobe's original scenario—specifically, that the event being evaluated was perceived as a side effect rather than as a means to an end. We were particularly concerned about the ECONOMIC-OUTLOOK case (we thank the Editor of this issue for raising this concern). To this end, we conducted a norming study on Prolific among neurotypical adults ($N = 39$) in which, after presenting Knobe's original vignette (i.e., the ENVIRONMENT vignette) as an example of a side effect and a variant of that vignette in which the pollution is a means to harm a rival company, participants were randomly shown 12 scenarios that the authors had classified as either means-to-an-end or side-effect. These scenarios included the STREET-FURNITURE and the ECONOMIC-OUTLOOK vignettes. The norming results showed an average misclassification rate of 24%—that is, 24% of the responses incorrectly identified a means-to-an-end event as a side-effect event, and vice versa. The error rates for the ECONOMIC-OUTLOOK and STREET-FURNITURE scenarios were 28% and 31%, respectively, with no significant differences when compared to the other scenarios.

We are surprised by the rate of incorrect responses across all the scenarios presented. It is therefore possible that, in general, judgments of intentionality are affected by poor comprehension of the vignettes[3]. However, we do not believe that this is an issue that specifically affects the two new scenarios we have introduced in this study. That said, as we mention below, we consider it possible that autistic participants had difficulty understanding that the event being evaluated in the ECONOMIC-OUTLOOK case was not a means-to-an-end, but a side-effect.

[3] It may be helpful to illustrate how the participants responded to the scenarios and what kind of scenarios they were asked to judge. For instance, participants were asked whether the flooding of a city described in the following vignette was a side effect or a means to some end: "The mayor of a city comes up with the idea of diverting the course of the river so that it flows through the center of the city. Someone warns him, 'The city will be beautiful, but if it rains a lot, there may be floods.' The mayor answers, 'It rains very little here, and the city is going to improve a lot.' The year after the works were completed, it rained like never before and the city flooded. The mayor was forced to resign". In this scenario, 23% of participants said that the flooding of the city was a means to some end, rather than a side effect of diverting the river.

3 Presentation of Key Findings

3.1 Attribution of Intentionality

We hypothesized that the asymmetry identified by Knobe (2003a) in the harm domain, and later reproduced by Zalla and Leboyer (2011) for autistic individuals and by Zucchelli and collaborators (2018) for individuals with autistic traits, would also be present in the aesthetic and alethic domains. The results confirmed this hypothesis, since for the three domains most participants attributed intentionality to the agent when the side-effect was negative, and a minor number of people attributed intentionality when the side-effect was positive.

As each participant was assigned to only one version of the scenario for each domain –either depicting a positive or negative side effect– the analyzed data represent two independent groups. A Pearson chi-square test of independence was conducted to examine whether the proportion of participants attributing intentionality differed between the positive and negative side-effect conditions. Since the analysis involved 2×2 contingency tables –with relatively small expected frequencies–, Yates' correction for continuity was applied to the chi-square tests to adjust for the potential overestimation of statistical significance (Yates 1934; Cochran 1954).

First, in the autistic group, we found that in the *harm domain* (i.e., ENVIRONMENT vignette), 92% of participants attributed intentionality to the agent in cases with a negative side effect, whereas only 13% made this attribution when the collateral effect was positive (see Fig. 1). This difference was highly significant ($\chi^2(1, N=99)=59.6, p<0.001$). These results are comparable to those of Zalla and Leboyer (2011), although the attributions of intentionality they obtained in scenarios with positive side effects were higher than those we found in these cases. In the case of neurotypical participants, we obtained rates of intentionality attribution –for the harm domain– very similar to those observed for autistic participants. More specifically, in the neurotypical group, 82% of participants attributed intentionality to the

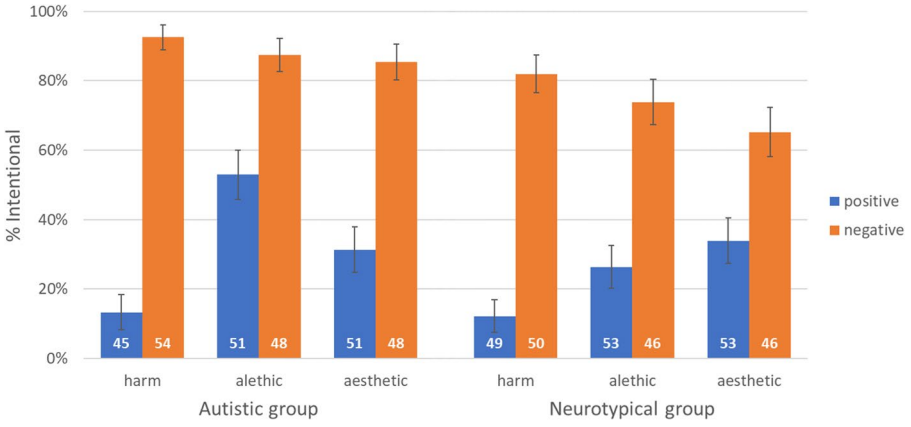

Fig. 1 Attribution of intentionality for the harm, aesthetic, and alethic domains. (Comparison between autistic vs. neurotypical groups.)

agent in cases with a negative side effect, whereas only 12% made this attribution when the side effect was positive. This difference was again highly significant ($\chi^2(1, N=99)=45.5, p<0.001$). These rates of intentionality attribution by the neurotypical group are very similar to those reported by Knobe (2003a); Zalla and Leboyer (2011).

In the aesthetic and alethic domains, there was also a significant asymmetry in the attribution of intentionality between the positive and negative cases in both populations (i.e., autistic and neurotypical), but some differences were found with respect to the harm domain. On the one hand, in the *aesthetic domain* (FURNITURE-URBAN vignette), 85% of autistic people attributed intentionality to the agent when the side effect of his action was ugly, whereas only 31% of them who attributed intentionality when the side effect was beautiful –a highly significant difference ($\chi^2(1, N=99)=27.4$, $p<0.001$), with an effect size was very close to that present in the harm domain. For neurotypicals, in the aesthetic domain, 65% of participants attributed intentionality to the agent when the aesthetic side effect of his action was ugly, whereas only 34% attributed intentionality when the side effect was beautiful. Again, the difference was significant ($\chi^2(1, N=99)=8.4, p=0.0037$).

On the other hand, in the case of the *alethic domain* (ECONOMIC-OUTLOOK vignette), the asymmetry was less pronounced in the autistic group, namely 87% of participants attributed intentionality to the agent when a false belief was produced as a side effect of his action, compared to 53% who attributed intentionality when the agent produced a true belief as a side effect. Although the size effect was smaller than that found in the harm and aesthetic domains, the difference was still highly significant ($\chi^2(1, N=99)=12.4, p<0.001$). In contrast, the asymmetry was more pronounced in the neurotypical group than in the autistic group (in fact, its size effect was quite close to that present in the harm domain). In particular, 74% of the neurotypical participants attributed intentionality to the agent when a false belief was produced as an alethic side effect, whereas only 26% attributed intentionality when a true belief was produced. This difference was again highly significant ($\chi^2(1, N=99)=20.4, p<0.001$).

The main differences between the autistic and neurotypical populations are found in two specific situations, where the intentionality attributed by the autistic group is significantly higher than that attributed by the neurotypical group: (a) when the side effect is negative in the aesthetic domain (85% for the autistic group vs. 65% for the neurotypical group, which is a significant difference, $\chi^2(1, N=94)=4.1, p=0.042$); and (b) when the side effect is positive in the alethic domain (53% for autistic vs. 26% for neurotypical, again a significant difference, $\chi^2(1, N=104)=6.6, p=0.01$). In addition, it is worth noting that the intentionality attributed by the autistic population is higher than that attributed by the neurotypical group in all cases with negative side effects.

As mentioned above, the effect size was also determined for the three domains using Cramer's V (VC). On the one hand, the effect size in the autistic group was large in both the harm (VC = 0.776) and aesthetic (VC = 0.526) domains, and medium in the alethic domain (VC = 0.354). On the other hand, in the neurotypical group it was found to be large in the harm domain (VC = 0.678), medium in the alethic domain (VC = 0.435), and small in the aesthetic domain (VC = 0.292). As expected, the largest difference between the autistic and neurotypical populations in terms of

effect size occurs in the aesthetic domain, with a large effect size for autistics and a small one for neurotypicals.

3.2 Attribution of Praise and Blame

In his original study, Knobe (2003a) also asked participants how much praise or blame the agent's action deserved. This allowed participants to decouple the attribution of intentionality from the moral blame that the agent's action deserved –insofar as they were allowed to attribute praise or blame to the agent before being asked to attribute intentionality to him. Similar questions were asked by Zalla and Leboyer (2011) in their study with autistic and neurotypical individuals.

In our case, we asked a similar initial question, with a twofold purpose: (1) to disentangle the attributions of intentionality and blame, and (2) to test for the presence of the moral component –through the attribution of praise or reprobation– in the aesthetic and alethic domains (see Fig. 2). As expected, the moral component is generally present in the aesthetic and alethic domains, as most of participants praise the agents for actions that end with a positive side effect and disapprove when their action results in a negative side effect. Furthermore, in cases with negative side effects, the greatest disapproval occurs in the harm domain (this was also expected for both negative and positive side effects), closely followed by the alethic domain. Similar results were obtained for the neurotypical group. However, certain differences emerged between the domains that deserve more detailed attention.

In the *harm domain*, 91% of autistic participants judged the chairman's action to be blameworthy in the case with a negative side effect, a percentage very close to the 94% of neurotypical participants who assigned blame in these situations. These rates of blame attribution for the negative cases are very similar to those reported by Zalla and Leboyer (2011). In contrast, in the cases with a positive harm side effect, the percentage of autistic participants attributing praise to the president decreased to 64%, and the

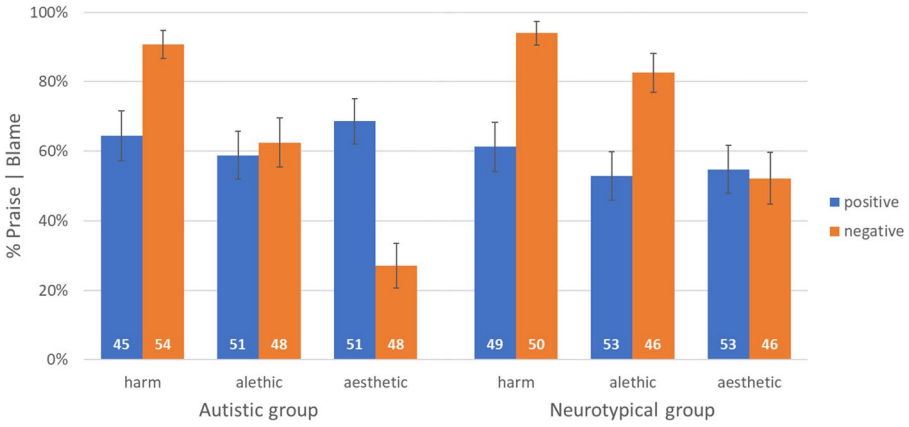

Fig. 2 Percentage of participants who judged the agent's actions to be praiseworthy or blameworthy (versus neutral) for positive and negative side effects, respectively. (Comparison between autistic vs. neurotypical groups.)

same happened to the attribution of praise by the neurotypical group, which dropped to 61%. As in the case of negative side effects, when the harm side effect is positive there are no significant differences between the percentages of autistic and neurotypical participants who judge the chairman's action as praiseworthy. However, on this point we find a difference with the results obtained by Zalla and Leboyer, since in their work the percentage of participants who attributed praise was lower for the neurotypical group (43%) and significantly lower for the autistic population (17%), which was a significant difference. This difference may be due to the much smaller sample size of Zalla and Leboyer's study ($N=46$ compared to 99 participants in our study).

On the other hand, in the *alethic domain*, the percentages of neurotypical participants attributing praise (53%) and blame (83%) to the agent (in the positive and negative cases, respectively) were significantly different ($\chi^2(1, N=99)=8.5$, $p=0.0035$), and quite similar to the rates found in the harm domain. However, the situation changes slightly for the autistic group, where, although the percentage of participants attributing praise (59%) is comparable to the same attribution by this population in the harm domain –and, also, to the percentage of neurotypical participants doing so in the alethic one–; in cases with a negative side effect, the percentage of autistic participants attributing blame decreases significantly (62%) with respect to the harm domain, with no significant difference between the attributions of praise and blame for the autistic group in the alethic domain ($\chi^2(1, N=99)=0.028$, $p=0.87$). Lastly, in the negative alethic side effect condition, the difference in blame attribution between the autistic group (62%) and the neurotypical group (83%) was significant ($\chi^2(1, N=94)=3.8$, $p=0.05$).

In relation to the *aesthetic domain*, 69% of the autistic participants judged the chairman's action to be praiseworthy in the case of a positive side effect, whereas only 27% judged his action as blameworthy in the case of a negative side effect, which is a highly significant difference ($\chi^2(1, N=99)=15.5$, $p<0.001$). In contrast, 55% of the neurotypical group attributed praise in the positive case, while 52% attributed blame in the negative case, which does not yield a statistically significant difference ($\chi^2(1, N=99)=0.0026$, $p=0.96$). Just as in the alethic domain, when there was a negative aesthetic side effect, the attribution of blame by the autistic population (27%) was significantly different from that of the neurotypical group (52%) ($\chi^2(1, N=94)=5.2$, $p=0.023$).

Based on the above, two clear patterns emerge in the attribution of blame or praise for the agent's actions. On the one hand, for the positive side effects, the percentage of participants attributing praise remains fairly constant at around 60% for both the autistic (~64%) and neurotypical (~56%) groups, and does not appear to be domain dependent –since there are no significant differences between any of them. On the other, for the negative side effects, there is a downward trend in the attribution of blame with domain, which is high for the harm domain, medium for the alethic domain, and low for the aesthetic domain. Differences in attributions of blame between domain pairs –for autistic and neurotypical populations– are highly significant for all except the harm vs. alethic comparison for neurotypical participants (see Table 1). Moreover, in the negative cases, the percentage of autistic participants who attribute blame is always lower than the percentage of neurotypical participants who do so.

Table 1 Differences in blame attributions between domain pairs for negative side effect cases

Group	Domain pair	χ^2	N	p-value	
Autistic	Harm (91%) vs. alethic (62%)	10.0	102	0.0015	**
	Harm (91%) vs. aesthetic (27%)	40.6	102	<0.001	***
	Alethic (62%) vs. aesthetic (27%)	10.8	96	0.001	***
Neurotypical	Harm (94%) vs. alethic (83%)	2.0	96	0.15	
	Harm (94%) vs. aesthetic (52%)	19.6	96	<0.001	***
	Alethic (83%) vs. aesthetic (52%)	8.4	92	0.0038	**
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

If we now represent the praise and blame attributed in the positive and negative side effect cases in the terms used by Knobe (2004), i.e., not only by examining whether participants judge the agent's behavior as praiseworthy or blameworthy –as Zalla and Leboyer (2011) do– but as the amount of praise and blame attributed to the agent in each case, the results still show the same two patterns mentioned above (see Table 2).

Therefore, it seems that there is a clear hierarchy of blame that would be common to both populations (i.e., autistic, and neurotypical), namely, the worst would be harming the environment, then inducing false beliefs in people, and finally causing aesthetic damage to the city. In addition, the neurotypical group always judges these cases more blameworthy than the autistic group.

4 Discussion of Results

4.1 Intentionality Attribution

The results of the experiment done with neurotypical participants confirm the findings of previous studies in the harm and aesthetic domains. Indeed, in the harm domain, the percentage of attribution of intentionality obtained was 82% for the negative side-effect and 12% for the positive side-effect, which is absolutely consistent with the results of the original work by Knobe (2003a) –in which the intentionality attributions obtained were 82% and 23% for the negative and positive side-effect

Table 2 Praise (+) and blame (−) attributed according to type of side-effect

Group	Domain	Positive effect	Negative effect
Autistic	Harm	1.71	−2.54
	Alethic	1.51	−1.88
	Aesthetic	1.84	−0.58
Neurotypical	Harm	1.63	−2.62
	Alethic	1.23	−2.11
	Aesthetic	1.42	−1.04

cases, respectively. Similarly, in the aesthetic domain, the intentionality attributions obtained were 65% and 34% for cases with negative and positive side-effects, respectively, which is also quite consistent with the results of Knobe (2004) for the variation of his original experiment with aesthetic damage (namely 54% and 18% for cases with negative and positive side-effects, respectively). On the other hand, the results in the alethic domain show an attribution of intentionality of 74% and 26% for cases with negative and positive collateral effects. This would confirm the presence of the asymmetry identified by Knobe, regardless of the type of side-effect present in the collateral effect (i.e., harm, aesthetic, or epistemic/alethic).

However, the effect size is not the same in these three domains. Knobe (2004) had already observed that the effect size in aesthetic evaluations was significantly smaller than the effect size in the harm cases. Our Experiment 1 confirms Knobe's observation, but also shows that, based on the size of the observed effect size, a gradation in the asymmetry of the intentionality attributed by the participants can be established as a function of the type of domain. This asymmetry would be large in the harm domain, medium in the alethic domain, and small in the aesthetic domain, and would correlate with the difference between the attributions of intentionality in the negative and positive cases in the harm (70%), alethic (47%), and aesthetic (31%) domains. This gradation of domains was confirmed by how neurotypical participants assigned blame to the agent's decision in cases with a negative side effect –i.e., harm (94%), alethic (83%) and aesthetic (52%). Therefore, the asymmetry in the attribution of intentionality is clearly domain dependent, and although this asymmetry was present in all three domains, we also found that the magnitude of the effect varied across them. A possible interpretation of these results is that participants judged that harming the environment was a worse outcome than misleading voters and that misleading voters was a worse outcome than spoiling the aesthetics of the city. The observed discrepancy between the harm scenario and the other two scenarios is likely attributable to the pivotal role that the concept of harm occupies within contemporary moral frameworks (Graham et al. 2009; Hanser 2019). Specifically, negative outcomes framed as instances of harm may tend to elicit stronger reactions than those framed otherwise, owing to the prominent societal emphasis currently placed on the moral dimension of harm.

With regard to this, Knobe's results in the aesthetic domain were that: "The mean rating for blame or praise in the aesthetic harm condition was −1.7; the mean in the aesthetic help condition was.3. This difference was statistically significant, $t(54) = 5.8$, $p < .001$" (Knobe 2004: 275). It was therefore interesting to see that in our case, the amount of blame attributed in cases with negative side effects was also greater than the amount of praise attributed in cases with positive side effects, while still observing the Knobe effect in intentionality attributions. This result supports Knobe's idea that intentionality judgments are triggered by consequences, regardless of how praiseworthy or blameworthy actions are judged.

In the autistic population, the Knobe effect was observed in the three domains (i.e., harm, aesthetic, and alethic). Interestingly, the effect size was smaller in the alethic case than in the other two domains, and was largest in the harm case.

4.2 Moral Evaluation

The above gradation is confirmed by the praise and blame that participants thought the agents' actions deserved. In regard with the *harm* domain, there are not many differences between autistic and non-autistic participants with respect to moral evaluations. Differences appear in the other two domains. Firstly, in the *alethic* domain, we observe a different pattern, i.e., the autistic evaluation is less negative than the neurotypical one in cases with a negative side effect, and quite similar in cases with positive side effects. Second, in the *aesthetic* domain we found the same pattern as in the alethic domain, only more pronounced, namely, the autistic evaluation is much less negative than the neurotypical in the negative cases, and not significantly different in the positive cases (i.e., when the city is aesthetically improved and looks more modern).

At first glance, the results could be seen as reflecting a profile of evaluations in the autistic case that is more consistent with a moral judgment based on the agent's declared intentions. In cases with bad consequences, on average, autistic people tend to consider agents less blameworthy than neurotypicals, as if they reasoned: "Knowing the side effect it could cause, such a side effect wasn't what the agent intended to do, so the agent should not be blamed for what happened." However, there is reason to suspect that this is not the case. On the one hand, when it comes to harming the environment, autistic participants do not seem to make their moral judgments independent of outcomes. On the other hand, if judgments were made on the basis of declared intentions, we should not expect differences between cases with good and bad outcomes. Nevertheless, in cases with good outcomes, autistic people give agents a level of praise very similar to that given by neurotypicals. Moreover, they praise the mayor in the aesthetic case slightly more than neurotypicals.

Taken together, the results suggest that autistic individuals are as consequentialist as neurotypicals in cases with positive side effects, but less consequentialist in cases with negative outcomes. This goes against the view that moral judgments in the autistic population are more outcome-based than neurotypical judgments due to theory of mind difficulties (Moran et al. 2011; Gleichgerrcht and Young 2013; Margoni and Surian 2016). Yet, it is difficult to hold that autistic individuals blame more on the basis of intentions than on the basis of outcomes, since, when it comes to intention attribution, we observe a strong Knobe effect in the autistic group. Moreover, as just mentioned, the non-consequentialist evaluative pattern is not stable, namely: it is not present in the harm case, and it cannot explain the difference between praise and blame observed in the aesthetic case.

Our results do not suggest a more ingrained sense of morality in autistic individuals, as we observe a large number of neutral evaluations. In this regard, it is interesting to note that neutral evaluations are more extended in the autistic group, which suggests that autistic people may be more ambivalent than neurotypical individuals especially in judging how blameworthy the agent's action is. This is consistent with evidence that higher levels of autism are associated with atypical patterns of moral judgment (Clarkson et al. 2023). The ambivalence we hypothesize may, in turn, be associated with heightened levels of general uncertainty (Van de Cruys et al. 2014; Bervoets et al. 2021). Such uncertainty may make participants provide a neutral judgement as a result of indecision.

4.3 General Comparison of Autistic vs. Non-Autistic Populations

The relationship between the intentionality attributions and the blame/praise judgments suggests that, in the case of autism, *intentionality judgments* are triggered by consequences, rather than by evaluations of the agent's actions. As noted above, this is consistent with Knobe's view and with our results with neurotypicals. The results with autistic participants in the aesthetic case dramatically illustrate this point.

When comparing neurotypical and autistic participants, we observe many similarities, and some differences. The most obvious similarity is the presence of the Knobe effect in all three domains. Another similarity between autistic and neurotypical populations worth mentioning is that both groups seemed to show a fairly stable hierarchy of evaluations by domain: (I) For negative effects, agents attributed more intentionality and blame in the harm domain, then in the alethic domain, and finally in the aesthetic domain. This occurred without exception, for both autistic and neurotypical groups, in all possible pairs of domains. (II) For positive effects, the amount of praise attributed is very similar in all cases (~60%), with no significant differences by domain or by group of participants.

Within this broad similarity, there are some differences. First, the degree of attribution of intentionality is higher in the autistic group overall, and it is higher in both cases: positive and negative. More specifically, it is particularly high (compared to the neurotypical group) in the aesthetic negative case and in the alethic positive case. The latter finding –i.e., the greater tendency of the autistic group to attribute intentionality, compared to the neurotypical population, when the side effect is positive in the alethic domain– may be explained by the theory of mind difficulties observed in autistic individuals (Baron-Cohen 2001; Garcia-Molina and Clemente-Estevan 2019).

In the case of blame/praise, autistic participants praise the agents more than neurotypicals in all three domains (i.e., harm, aesthetic, and alethic) –although these differences were not significant– and blame them less in the three domains as well, especially in the alethic and aesthetic domains, thus exhibiting a greater difference between intention and blame/praise judgments in these cases. Finally, in the harm domain, the results of autistic participants are very similar to those of neurotypicals for both types of judgments. This is in partial contrast to the results of Zalla and Leboyer (2011), who found a significant difference for praise judgments.

Overall, autistic participants seem to exhibit a stronger Knobe effect than neurotypicals, but also a greater degree of intentionality attribution (in both positive and negative outcomes). However, the difference in the case of positive outcomes case is mainly due to the alethic domain. Recall that in this scenario the reader is confronted with an agent who, as a side effect of her action, either misleads her audience or lets them know the truth. In the latter case, autistic participants are much more likely than neurotypicals to believe that the candidate is intentionally letting the audience know the truth. This scenario involves more mindreading than the other two, as it involves not only tracking the agent's behavior, but also its effect on the agent's audience. That is, it is a communication scenario that may be difficult for autistic people to fully grasp. In particular, it may be unclear to them from reading the vignette to what extent the candidate really intended to tell the truth or not. The fact that their praise

and blame judgments are similar also suggests some difficulty in understanding the scenario, as they do not seem to blame the intentional deceiver much more than they praise the one who ends up providing accurate information to her audience without caring about the truth. If the intergroup differences in the alethic case are indeed related to its intricate theory of mind narrative, the results of the other scenarios strongly suggest a larger Knobe effect in the autistic population.

Regarding praise/blame evaluations, the harm case is almost identical in the autistic and neurotypical populations. The observed differences concern the other two scenarios. Leaving aside the alethic scenario and focusing on the aesthetic case, we see that the pattern observed in the neurotypical group also appears and is accentuated in the autistic group. The aesthetic praise/blame results are difficult to interpret in the light of previous Knobe effect studies. Participants are unclear about the trade-off between making a city ugly and making it look more modern. Obviously, they think that it is good to make the city both more beautiful and more modern, but they may think that it is not so bad to make it more modern but uglier. However, since the result is negative (not so bad, but still negative), the Knobe effect is triggered. Therefore, we see the following pattern: a clear Knobe effect combined with more praise than blame, which is more pronounced in the autistic population. That is, autistic individuals seem to think that the wrong that the mayor brought about is permissible but at the same time that he intended it. This suggests that the point Knobe often makes, i.e., that judgments of intention are divorced from moral judgments, is even more acute in the case of autistic individuals. This is consistent with Zalla & Leboyer's findings for autistic individuals, although we would not reach this conclusion if we focused only on the harm case, which is the one they studied.

We expected to observe higher moral standards in the praise/blame questions in the autistic group, based on the idea that autistic people may be stricter with respect to morals than neurotypicals. However, this difference did not show up in the results. The prediction that was most borne out was the one about the Knobe effect, namely, that we would observe a larger asymmetry in the attribution of intentionality for the positive and negative cases –and thus a larger Knobe effect– in the autistic group than in the neurotypical group. This was confirmed by higher intentionality attributions in all three domains for the negative cases, and lower for the positive cases (except for the alethic domain), leading to higher effect sizes in most domains (i.e., moral and aesthetic). We hypothesized that such results could relate to difficulties distinguishing between incidental and intentional outcomes, and some persistent differences in means-ends evaluations.

4.4 Gender Differences

We conducted a post-hoc analysis based on the gender variable, since we wanted to know whether there could be differences between males and females, particularly in the autistic group, given that many authors have recently emphasized that the female autistic profile may differ in some ways from the male autistic profile (e.g., Kirkovski et al. 2013).

The distribution of males and females in the autistic and neurotypical groups was as follows. In the neurotypical group, 65% of the subjects were female ($N = 59$)

and 35% were male ($N=32$) –excluding 8 subjects who answered "other" or did not respond to the gender question. In the autistic group, the distribution between males ($N=47$) and females ($N=48$) was more balanced, with each group representing approximately 50% of the total –after excluding 4 subjects who answered "other" to the gender question.

Quite surprisingly, some important gender differences were found in *both kinds of populations*. Concerning intentionality attribution (see Fig. 3). In the first place, neurotypical males attributed very little intentionality (27%) to cases with negative side effects in the aesthetic domain, to the extent that no Knobe effect was observed in this domain, whereas neurotypical females displayed a more consistent pattern of intentionality attributions across domains, with a strong Knobe effect in all cases. In second place, autistic males attributed a lot of intentionality to cases with positive side effects in the alethic domain (68%) –a case in which Knobe's asymmetry did not occur– as well as in the aesthetic domain (46%). In contrast, the profile of responses of autistic females was again more consistent across domains, being very similar to that of non-autistic females.

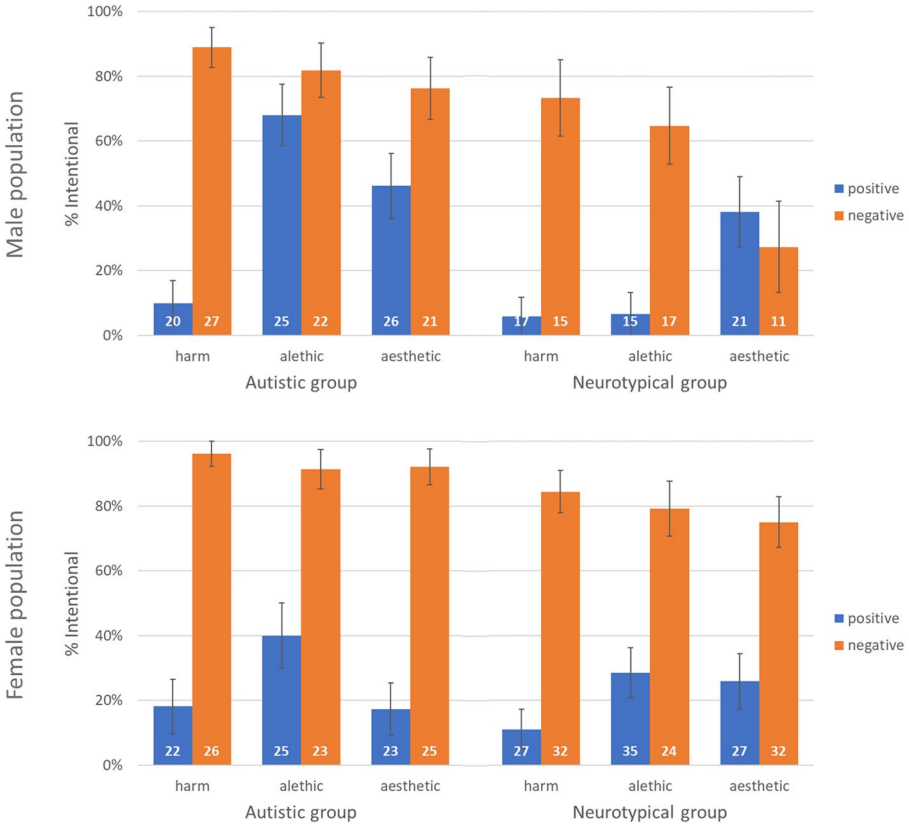

Fig. 3 Attribution of intentionality for the harm, aesthetic, and alethic domains. (Comparison by gender in autistic and neurotypical groups.)

These general results suggest that the interdomain differences found at the general level regarding intentionality attribution are attributable to males. In the case of the aesthetic domain, it appears that neurotypical males might show a low sensitivity to damage in the aesthetics of the city (i.e., they may not consider the side-effect to be as bad). More interesting, we think, is the case of the alethic domain within the autistic population. Above we mentioned that the attenuated Knobe effect found in the autistic group could relate to theory of mind difficulties: some autistic individuals might have had difficulty understanding that the politician in the vignette was not concerned with whether potential voters would know the truth or not. If this is the explanation for the observed differences, it seems to affect mainly autistic males.

Regarding moral judgments, relevant gender differences were also observed (see Fig. 4). First, neurotypical women attributed praise and blame in the alethic domain in exactly the same way as they did in the harm domain, whereas neurotypical men showed a much more graded scale in the attribution of blame. This suggests that neurotypical females tend to think that deceiving people is as morally wrong as harming

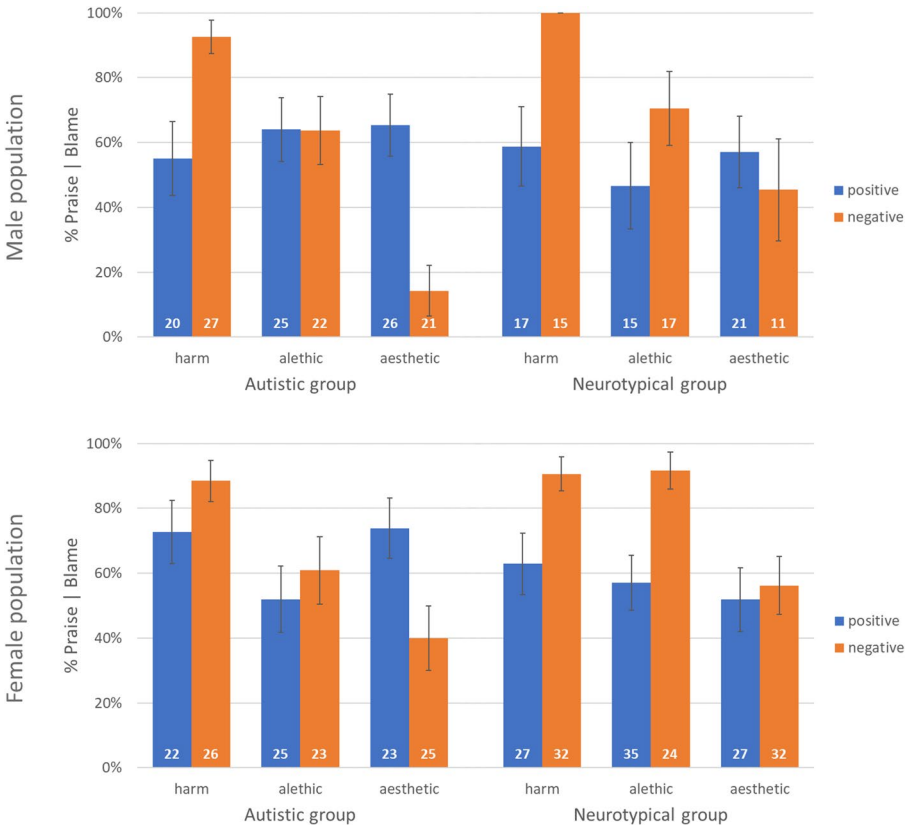

Fig. 4 Percentage of participants who judged the agent's actions to be praiseworthy or blameworthy (versus neutral) for positive and negative side effects, respectively. (Comparison by gender in autistic and neurotypical groups.)

the environment. This difference with respect to the alethic case was also observed, although attenuated, in the autistic group.

Specifically for the autistic group, the most striking finding was that autistic males did not assign blame for actions that caused harm in the aesthetic domain, although they believed that the agent had intentionally provoked the side effect. As in the case of intentionality attribution, autistic females displayed a more consistent pattern of evaluations overall than autistic males. However, they did not differ much from non-autistic females.

Taken together, these results suggest that, with the exception of the stronger Knobe effect in autism, the intergroup differences reported above concerning intentionality attribution, moral judgment, and the detachment of intentionality attribution from moral evaluation are primarily attributable to differences in the male populations.

4.5 Back To the Knobe Effect

Finally, of all the explanations of the Knobe effect on the market, the normativist interpretation is perhaps the one that can best explain the results obtained. According to the normativist explanation, the asymmetric attribution of intentionality by participants is due to the fact that people form stronger beliefs about collateral effects in negative cases –where a norm violation occurs– than in positive cases –or norm-compliant cases–, and therefore the attribution of intentionality would be higher in the former (Holton 2010; Alfano et al. 2012). The normativist explanation could explain the decreasing rates of intentionality and blame attribution observed in cases with negative collateral effects across domains (both overall and when results were analyzed by gender). The highest rates were observed in the harm domain, where the side effect can be clearly identified as a norm violation of the environmental protection laws. Then, in the alethic domain, the attributions of intentionality and blame would be lower because the side effect –i.e., the induction of wrong beliefs– was less clearly associated with a norm violation. The lowest attributions of intentionality and blame were found in the aesthetic domain, where the negative side effect was much more difficult to identify as a norm violation.

This gradient in the attribution of intentionality –and blame–, from harm to alethic to aesthetic domains, is consistent with the normativist account by illustrating how the perceived violation of norms influences the attribution of intentionality. In domains where the negative collateral effects are more clearly seen as norm violations, intentionality is attributed to a greater extent. Conversely, in domains where the association with norm violation is less clear, the attribution of intentionality decreases. This pattern highlights the impact of normativist considerations on people's judgments of intentionality.

5 Conclusions

The present study of the presence of the Knobe effect in the harm, alethic, and aesthetic domains –for both autistic and neurotypical populations–, has led to some relevant findings, which in several cases go beyond the results of previous studies.

First, the responses of the neurotypical population exhibited a high degree of consistency with the widely replicated result in the harm domain –i.e., an asymmetry in intentionality attribution consistent with the Knobe effect– and confirmed that such an effect is also present in the aesthetic and alethic domains. Moreover, our study has shown that this asymmetry is not only present in these three domains, but also that its effect size varies across domains, being large in the harm domain, somewhat smaller in the alethic domain, and even smaller in the aesthetic domain. The above gradation of effect sizes across domains provides a nuanced view of how individuals evaluate the outcomes of actions, insofar as the perceived harm, epistemic, or aesthetic consequences of an action affect the attribution of intentionality differently, with harm consequences having the strongest effect.

Second, in the case of the autistic population we have also found the presence of the Knobe effect in the harm, alethic, and aesthetic domains, although autistic participants showed a stronger Knobe effect than neurotypicals. Furthermore, the gradation between domains identified in the neurotypical group was also identified in the autistic population. The gradation was confirmed by the praise/blame that both autistic and neurotypical participants attributed to the agent's actions when the side effects were negative.

Overall, our results support Knobe's original idea that attributions of intentionality are triggered by the outcomes of actions. Compared to the neurotypical population, our results suggest that autistic individuals are just as consequentialist for positive outcomes, but less consequentialist for negative outcomes, and their intentionality judgments would be driven primarily by consequences rather than by evaluations of the agent's actions. In addition, our results also suggest that the autistic population may be more ambivalent than neurotypicals in attributing blame. All of this confirms Knobe's point that judgments of intentionality are divorced from moral judgments, which, in line with previous work in the harm domain, would be even more true for the autistic population. Yet, after analyzing the results by gender, we found that most of the observed intergroup differences were related to differences in the responses of males. Autistic and non-autistic females showed very similar response patterns. The only intergroup difference that remains is that the autistic group exhibits a stronger Knobe effect than the non-autistic group.

Lastly, of all the explanations for the Knobe effect on the table, our results are more consistent with the normativist interpretation, according to which the greater intentionality attributed in the negative cases is due to the fact that they are instances of violations of norms. This interpretation would explain the larger effect size in both the harm and the alethic domains, and the smaller effect size in the aesthetic domain. The fact that our findings support the normativist thesis suggests that further empirical research is needed to elucidate the nature and origin of the Knobe effect, in both autistic and neurotypical populations. Such studies would help to determine how perceptions of norm violation may affect judgments of intentionality and morality, and whether this influence may vary significantly as a function of social, cultural, or economic factors.

Appendix 1. Distractor Scenarios

This appendix presents the six distractor scenarios employed in the experiment. Each distractor scenario was designed to be similar in form to the critical scenarios, yet without the collateral effect characteristic of cases involving the Knobe effect. Detailed descriptions of the six distractor scenarios are provided below.

Distractor scenario I: The personnel manager of a company went to the human resources director and said, "We have designed a new program to promote work-life balance for our employees, but it will have a significant internal cost to the company." The human resources director answered, "I do not care at all about the cost this has to the company. I just want the employees to strike a balance between work and family. Let's start the new program." The company started the new program. Sure enough, it had a significant cost to the company.

And then the participants were asked:

> [1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:
>
> The human resources director did the right thing by starting the new program.
>
> [2] Did the human resources director intentionally contribute to the employees having more time to take care of their families?

Distractor scenario II: The head of sales of a company went to the personnel director and said, "We have designed a new program to reorganize the sales force. It will reduce the number of sales managers, and it will also worsen the communication processes." The director answered, "I do not care at all about worsening the communication processes. I just want to reduce the number of managers in the sales network. Let's start the new program." The company started the new program. Sure enough, the communication worsened.

And then the participants were asked:

> [1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:
>
> The personnel director did the right thing by starting the new program.
>
> [2] Did the personnel director intentionally contribute to the dismissal of the sales managers?

Distractor scenario III: The museum and library coordinator of a city council went to the city councilor for culture and said, "We have designed a program to update the library materials of the local libraries, but as a result, we will not have a budget surplus." The councilor for culture answered, "I do not care at all about not

having a budget surplus this year. I just want to expand the catalog of our libraries. Let's start this program." The city council started the program. Sure enough, they did not have a budget surplus that year.

And then the participants were asked:

[1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:

The councilor for culture did the right thing by starting the new program.

[2] Did the councilor for culture intentionally contribute to making a larger library catalog available to the public?

Distractor scenario IV: The executive deputy secretary for water and environmental protection went to the Secretary of State for Agriculture and said, "The selection process for the sewage treatment is over and one of the proposals was sent by your brother's company. Accepting it would damage the image of the selection process." The Secretary of State for Agriculture answered, "I do not care at all about damaging the image of the selection process. I just want the proposal submitted by my brother's company to be considered. Let's accept all the submitted proposals." All the submitted proposals were accepted. Sure enough, this damaged the image of the selection process.

And then the participants were asked:

[1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:

The Secretary of State for Agriculture did the right thing by accepting all of the submitted proposals.

[2] Did the Secretary of State for Agriculture intentionally favor the proposal sent by his brother's company?

Distractor scenario V: The editor-in-chief of a newspaper went to the editor and said, "We are thinking of publishing this news story that we have been working on for several weeks. However, our advertisers might not like it." The newspaper editor answered, "I do not care at all about how the advertisers react to the piece of news. I just want it to be known by the public opinion. Let's publish it on the front cover of tomorrow's edition." The piece of news was published on the front cover the next day. Sure enough, the publication annoyed some of the newspaper's advertisers.

And then the participants were asked:

[1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:

The newspaper editor did the right thing by publishing the piece of news.

[2] Did the newspaper editor intentionally contribute to his readers being well-informed??

Distractor scenario VI: The employment coordinator of a university went to the vice-president of human resources and said, "We are thinking of inviting the leading companies in the region to the career fair. However, if most of them accept the invitation, it will be difficult to find a place to accommodate all of them properly." The vice-president of human resources answered, "I do not care at all about reducing the space given to each company in the career fair. I just want the number of participating companies to be as large as possible. Let's send an invitation to all the mentioned companies." The leading companies in the region were invited. Most of them attended the career fair. Sure enough, there was not enough space to accommodate them properly.

And then the participants were asked:

[1] Rate your degree of agreement (+3) or disagreement (–3) with the following statement:

The vice-president of human resources did the right thing by planning the event that way.

[2] Did the vice-president of human resources intentionally contribute to maximizing the number of companies that could contact the students at this event?

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13164-025-00792-x.

Acknowledgements We are very grateful to two anonymous reviewers and the editor of this journal, whose comments and suggestions helped us improve the paper. We also want to thank Elena Castroviejo, Irene García Molina, Nuria Esteve, Valentina Petrolini and Greta Mazzaggio for their feedback.

Authors' contributions José V. Hernández-Conde: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing, Project Administration, Funding Acquisition; Agustín Vicente: Conceptualization, Investigation, Writing – Original Draft, Writing – Review, Funding Acquisition.

Funding Open access funding provided by FEDER European Funds and the Junta de Castilla y León under the Research and Innovation Strategy for Smart Specialization (RIS3) of Castilla y León 2021-2027. All authors were supported by the Agencia Estatal de Investigación and Ministry of Science and Innovation (grant number PID2021-122233OB-I00) and a BBVA Foundation Grant for Scientific Research Projects 2021 (RILITEA). The Foundation takes no responsibility for the opinions, statements and contents of this project, which are entirely the responsibility of its authors. JH's research was supported by a 2022 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the author.

Data Availability Data available on request from the authors.

Declaration

Ethical approval Ethical approval was issued by the Ethics Committee for Research Involving Medicinal Products in Valladolid Health Area (CEIm Área de Salud Valladolid Este) --responsible of ethical approvals for the University of Valladolid in 2022--, code PI_22-2713_NO_HCUV.

Consent Prior to enrollment, each potential participant received a written informed-consent form that clearly outlined the study's objectives, procedure, and expected duration, along with details on data confidentiality and how results would be used. The form stressed that taking part was entirely voluntary. It also underscored the right of revocation: participants could withdraw their consent and discontinue involvement at any moment, and they could request that any data already collected from them be excluded from the analysis without needing to provide a reason.

Competing interests The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Knobe, J. 2003b. Intentional action in folk psychology: An empirical investigation. *Philosophical Psychology* 16 (2): 309–324.

Knobe, J. 2004. Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology* 24 (2): 270–279.

Adams, F. 2004b. Intentional actions and moral considerations: Still pragmatic. *Analysis* 64 (3): 264–267.

Knobe, J. 2022. Morality and possibility. In *The Oxford handbook of moral psychology*, ed. M. Varga, and J. M. Doris. 310–332. Oxford: Oxford University Press.

Margoni, F., and Surian, L. 2022. Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology* 41(12):8774–8783.

Adams, F., and A. Steadman. 2004a. Intentional action in ordinary language: Core concept or pragmatic understanding. *Analysis* 64 (2): 173–181.

Alfano, M., J.R. Beebe, and B. Robinson. 2012. The centrality of belief and reflection in Knobe-effect cases: A unified account of the data. *The Monist* 95 (2): 264–289.

Baron-Cohen, S. 2001. Theory of Mind and autism: A review. In *International review of research in mental retardation: Autism*, vol. 23, ed. L.M. Glidden, 169–184. New York: Academic.

Bervoets, J., D. Milton, and S. Van de Cruys. 2021. Autism and intolerance of uncertainty: An ill-fitting pair. *Trends in Cognitive Sciences* 25(12):1009–1010.

Blair, R.J.R. 1996. Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders* 26 (5): 571–579.

Clarkson, E., J.D. Jasper, J.P. Rose, G.J. Gaeth, and I.P. Levin. 2023. Increased levels of autistic traits are associated with atypical moral judgments. *Acta Psychologica* 235: 103895.

Cochran, W.G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* 10 (4): 417–451.

Cova, F. 2015. The folk concept of intentional action: Empirical approaches. In *A companion to experimental philosophy*, ed. L. Sytsma, 121–141. Oxford: Blackwell.

Cova, F., A. Lantian, and J. Boudesseul. 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* 42 (10): 1295–1308.

Cushman, F. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108 (2): 353–380.

Cushman, F., and A. Mele. 2008. Intentional action: Two-and-a-half folk concepts? In *Experimental philosophy*, ed. J. Knobe and S. Nichols, 171–188. Oxford: Oxford University Press.

Dalbauer, N., and A. Hergovich. 2013. Is what is worse more likely? The probabilistic explanation of the epistemic side-effect effect. *Review of Philosophy and Psychology* 4 (4): 639–657.

Dempsey, E.E., C. Moore, A. Johnson, S.H. Stewart, and I.M. Smith. 2020. Morality in autism spectrum disorder: A systematic review. *Development and Psychopathology* 32 (3): 1069–1085.

Feltz, A., and E.T. Cokely. 2024. Intentions and side effects. In *Diversity and disagreement: From fundamental biases to ethical interactions*, 61–102. Cham: Palgrave Macmillan.

Garcia-Molina, I., and R. A. Clemente-Estevan. 2019. Autism and faux pas. Influences of presentation modality and working memory. *The Spanish Journal of Psychology* 22, e13: 1–11.

Gleichgerrcht, E., and L. Young. 2013. Low levels of empathic concern predict utilitarian moral judgment. *PLoS One* 8 (4): e60418.

Graham, J., J. Haidt, and B.A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96 (5): 1029–1046.

Graham, J., J. Haidt, M. Motyl, P. Meindl, C. Iskiwitch, and M. Mooijman. 2018. Moral foundations theory: On the advantages of moral pluralism over moral monism. In *Atlas of moral psychology*, ed. K. Gray and J. Graham, 211–222. New York: The Guilford Press.

Guglielmo, S., and B.F. Malle. 2010. Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin* 36 (12): 1635–1647.

Hanser, M. 2019. Understanding harm and its moral significance. *Ethical Theory and Moral Practice* 22 (4): 853–870.

Happe, F., and U. Frith. 2006. The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 36 (1): 5–25.

Hill, E. L. 2004. Executive dysfunction in autism. *Trend in Cognitive Sciences* 8(1):25–32.

Holton, R. 2010. Norms and the Knobe effect. *Analysis* 79(3):417–424.

Hu, Y., A.M. Pereira, X. Gao, B.M. Campos, E. Derrington, B. Corgnet, X. Zhou, F. Cendes, and J.-C. Dreher. 2021. Right temporoparietal junction underlies avoidance of moral transgression in autism spectrum disorder. *The Journal of Neuroscience* 41 (8): 1699–1715.

Kirfel, L., and J. Phillips. 2023. The pervasive impact of ignorance. *Cognition* 231: 105316.

Kirkovski, M., P.G. Enticott, and P.B. Fitzgerald. 2013. A review of the role of female gender in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 43 (11): 2584–2603.

Knobe, J. 2003a. Intentional action and side effects in ordinary language. *Analysis* 63 (3): 190–194.

Laurent, S.M., B.J. Reich, and J.L.M. Skorinko. 2021. Understanding side-effect intentionality asymmetries: Meaning, morality, or attitudes and defaults? *Personality and Social Psychology Bulletin* 47 (3): 410–425.

Machery, E., and T. Zalla. 2014. The concept of intentional action in high-functioning autism. In *Oxford studies in experimental philosophy: Volume 1*, ed. J. Knobe, T. Lombrozo, and S. Nichols. 152–172. Oxford: Oxford Academic.

Margoni, F., and L. Surian. 2016. Mental state understanding and moral judgment in children with autistic spectrum disorder. *Frontiers in Psychology* 7: 1478.

Moran, J.M., L.L. Young, R. Saxe, S.M. Lee, D. O'Young, P.L. Mavros, and J.D. Gabrieli. 2011. Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America* 108 (7): 2688–2692.

Nadelhoffer, T. 2004. On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology* 24 (2): 196–213.

Nakamura, K. 2018. Harming is more intentional than helping because it is more probable: The underlying influence of probability on the knobe effect. *Journal of Cognitive Psychology* 30 (2): 129–137.

Ngo, L., M. Kelly, C.G. Coutlee, R.M. Carter, W. Sinnott-Armstrong, and S.A. Huettel. 2015. Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports* 5: 17390.

Nichols, S., and J. Ulatowski. 2007. Intuitions and individual differences: The knobe effect revisited. *Mind and Language* 22 (4): 346–365.

Pettit, D., and J. Knobe. 2009. The pervasive impact of moral judgment. *Mind and Language* 24 (5): 586–604.

Phillips, J., J.B. Luguri, and J. Knobe. 2015. Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition* 145: 30–42.

Shulman, C., A. Guberman, and N. Shiling. 2012. Moral and social reasoning in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 42 (7): 1364–1376.

Stewart, S.L.K., B.J. Kennedy, and M. Haigh. 2022. Valence of agents and recipients moderates the side-effect effect: Two within-subjects, multi-item conceptual replications. *Journal of Cognitive Psychology* 34 (2): 289–306.

Strang, J.F., L.G. Anthony, B.E. Yerys, K.K. Hardy, G.L. Wallace, A.C. Armour, and L. Kenworthy. 2017. The flexibility scale: Development and preliminary validation of a cognitive flexibility measure in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 47 (8): 2502–2518.

Uttich, K., and T. Lombrozo. 2010. Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition* 116 (1): 87–100.

Van de Cruys, S., K. Evers, R. Van der Hallen, L. Van Eylen, B. Boets, L. de-Wit, and J. Wagemans. 2014. Precise Minds in uncertain worlds: Predictive coding in autism. *Psychological Review* 121(4):649–675.

Yates, F. 1934. Contingency tables involving small numbers and the χ^2 test. *Supplement To the Journal of the Royal Statistical Society* 1(2):217–223.

Young, L., F. Cushman, R. Adolphs, D. Tranel, and M. Hauser. 2006. Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture* 6 (1–2): 265–278.

Zalla, T., and M. Leboyer. 2011. Judgement of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology* 2:681–698.

Zucchelli, M.M., R. Nori, E. Gambetti, and F. Giusberti. 2018. The influence of high autistic personality traits on the attribution of intentionality in typically developing individuals. *Journal of Cognitive Psychology* 30 (8): 840–853.

Zucchelli, M.M., F. Starita, C. Bertini, F. Giusberti, and E. Ciaramelli. 2019. Intentionality attribution and emotion: The Knobe effect in alexithymia. *Cognition* 191: 103978.

Zucchelli, M. M., Matteucci Armandi Avogli Trotti, N. Pavan, A. Piccardi, and L. Nori, R. 2025. The dual process model: The effect of cognitive load on the ascription of intentionality. *Frontiers in Psychology* 16:1451590.