AI & SOCIETY
https://doi.org/10.1007/s00146-024-02058-z

OPEN FORUM

Explainable artificial intelligence and the social sciences: a plea for interdisciplinary research

Wim De Mulder[1]

Received: 19 January 2024 / Accepted: 5 August 2024

Abstract
Recent research emphasizes the complexity of providing useful explanations of computer-generated output. In developing an explanation-generating tool, the computer scientist should take a user-centered perspective, while taking into account the user's susceptibility to certain biases. The purpose of this paper is to expand the research results on explainability from the social sciences, and to indicate how these results are relevant to the field of XAI. This is done through the presentation of two surveys to university students. The analysis of the results leads to some interesting hypotheses, for example that the presented order of historical facts might be more influential on the interpretation or appreciation of an event than the actual temporal order of these facts. The computer scientist should, therefore, pay particular emphasis to the format of the produced output of any explainable artificial intelligence system. The main message of the paper is that results from the social sciences must be regarded as a crucial foundation of any explainable artificial intelligence system.

Keywords Explainability · XAI · User-centered perspective · Counterfactuals

1 Introduction and purpose of the paper

Whereas in the past the use of an AI system was mainly judged in terms of accuracy and computation time, other parameters have come to play a role in the appreciation of an intelligent system. The evaluation measure that has received the most attention in recent years is the degree of transparency (Adadi and Berrada 2018). In particular in fields where high-stakes decisions are involved, it is argued that AI models must provide explanations that reveal their inner workings (Rudin and Radin 2019). Explainable AI (XAI) is supposed to be beneficial for service providers (e.g. executives in lending companies), since it increases trust in the system, as well as for affected individuals (e.g. potential borrowers), as motivated decisions ensure better acceptance by these individuals (Sachan et al. 2020).

Recently, it has become clear that the provided explanations should not only depend on the modeled phenomenon but also on the wider context. In particular, different categories of users (e.g. expert versus novice users) might require different types of explanations (Wang et al. 2019). Yet, the computer scientist who developed the model might not be in the best position to implement an explanation generating tool. Miller rightly criticizes the fact that research in XAI typically does not cite or build on frameworks of explanation from the social sciences; rather, computer scientists rely on their intuition of what constitutes a 'good' explanation (Miller 2019). For example, while it might be intuitive that the user is optimally informed when a description of the complete chain of causes of a certain event is provided, research from the social sciences shows that people prefer a very limited number of causes, even if these causes only partially explain the event (Trabasso and Bartolone 2003).

The purpose of this paper is to expand the research results on explainability from the social sciences, but with a particular focus on contributing to the field of XAI. To this end, we constructed two fictitious but realistic cases, and we asked participants to rate to what extent certain, possibly hypothetical, information is or might be useful in explaining a given event. The first case involves the bankruptcy of an enterprise (relating to the field of the economy), while

This work was supported by the Research Foundation - Flanders (Grant number G006421N).

✉ Wim De Mulder
wim.demulder@kuleuven.be

[1] Centre for IT & IP Law, KU Leuven, Sint-Michielsstraat 6 - box 3443, 3000 Leuven, Belgium

Published online: 05 September 2024

the second case describes a criminal offence (relating to the domain of law). These cases were chosen to extend the usual layman cases encountered in social science research (e.g. about a person who arrived late home from work to find his wife unconscious on the floor Girotto et al. 1991) to cases that require expert knowledge to fully understand all details. Such cases better suit the goal of developing explainable computer models than the surveys developed in conventional social science research, since computer systems are typically being developed for specialized purposes (Thompson and Spanuth 2018). The cases were presented to students at universities in Belgium, mainly from the Faculty of Law. The main reason for this selection of participants is that it allows us to make a comparison between the two cases in terms of background knowledge, i.e. do the responses differ between the case where some expert knowledge is present (criminal offence case) and the case where less background knowledge is present (bankruptcy case)?

The outline of the paper is as follows. In Sect. 2 we describe how the survey was created. Details are provided on the design of the survey questions, on the participants, and on the method of analysis of the responses. In Sect. 3 we describe the first case along general lines, including the questions that were presented to the participants, and we provide an analysis of the responses. The same is done for the second case in Sect. 4. An analysis at a higher level, where the results of both cases are integrated and compared, is provided in Sect. 6 as a conclusion. The full details of the cases, as they were presented to the respondents, are presented in the Appendix.

2 Methods

2.1 Research questions

The overall research question is what kind of explanations are required by users to understand a given event. The survey questions have been developed with the particular intention of contributing to the field of XAI. In this respect most of our research questions are related to some major findings from the social science literature that are relevant to the field of XAI, some of them outlined in the aforementioned work by Miller (2019). In particular, the following research questions are addressed (for each research question the section is mentioned where the research question is outlined in more detail):

- Q1: To what extent are counterfactuals appreciated as explanations of a particular event (Sect. 3.1.4)?
- Q2: Do users with background knowledge in a particular field prefer specialized explanations over explanations that are easily understood by the wider public (Sect. 4.1.3)?
- Q3: Are users able to detect irrelevant information, and subsequently discard it as a non-explanation, in particular when this information concerns the moral character of a human being (Sect. 3.1.2)?
- Q4: To what extent do users consider suggestive but non-factual statements as explanations (Sect. 4.1.2)?
- Q5: Does the degree of explainability of multiple explanations depend on the order in which these explanations are presented (Sect. 3.1.3)?

The last research question has, as far as we are aware, not yet been considered. Miller refers to social science research results that show that the temporality of events has an impact on the degree of explainability as assumed by people (Miller and Gunasegaram 1990), but this leaves open the question whether it is the temporal order of the events itself *or* the temporal order of *the presented facts* that is of main relevance.

2.2 Research design

The survey was constructed with the software tool Qualtrics.[1] Participants were randomly presented with one of several variants of each of both cases (the cases are described below). The variants are only slightly different (in terms of the contents of the case), and the variants were so created that a comparison between the responses to the variants enables analysis of the above research questions. After having been asked to read carefully the given case, respondents were presented with several statements, and they were invited to rate the degree to which they agree with it on a 5-point Likert scale (Joshi et al. 2015). There is debate on the optimal number of scale points (Chang 1994), but we found that more than 5 scale points resulted in a cluttered layout. The 5 scales were as follows: "strongly disagree", "slightly disagree", "no opinion", "slightly agree" and "strongly agree". All statements relate, implicitly or explicitly, to the concept of explainability. In case a group of statements was presented, the statements were given in a random order to control for confounding in the analyses of the results.

2.3 Participants

Students at the Faculty of Engineering, Law, and Economics at Ghent University were invited, through e-mail, to participate in the survey. This selection was simply motivated by the fact that the survey is part of an interdisciplinary research project at these faculties, which implied easy access

[1] https://www.qualtrics.com.

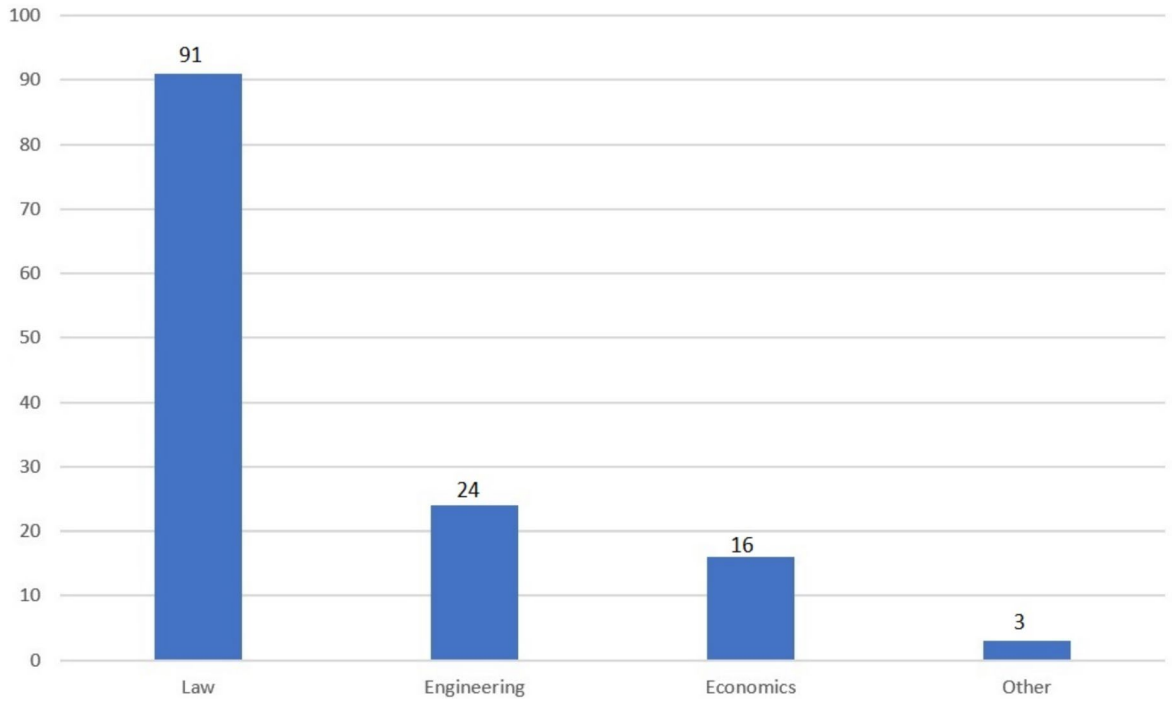


Fig. 1 Overview of the respondents

Table 1 Number of effective respondents for the first case

Variant	Number of effective respondents
1	34
2	39
3	39

Table 2 Number of effective respondents for the second case

Variant	Number of effective respondents
1	51
2	48

to these students by the project members. Students were first asked to indicate to which faculty they belong, where the option 'Other' was provided since some students only take up a very limited number of courses at one of the aforementioned faculties while mainly belonging to another faculty. There were 134 participants, the vast majority of whom were law students, cf. Fig. 1.

However, some respondents only participated *pro forma*, in the sense that they only provided their personal details, but did not answer any of the questions related to the cases. Tables 1 and 2 show the number of effective respondents for both cases, i.e. the respondents that were take into account for analysis.

The fact that there is a mix of law students and other students, might have an impact on the results of the second case. The reason is that this case refers to the legal field, implying that respondents with background knowledge (law students) might perceive the given case differently than respondents who lack such knowledge (other students). Therefore, for this case we provide a twofold analysis: one where all respondents are taken into account, and one where only responses from law students are taken into account.

2.4 Analysis of the results

Responses were first analyzed by visual comparison of the bar charts of responses for the different variants. Next, the Kruskal–Wallis test (Kruskal and Wallis 1952) was applied to identify statistically significant differences between the variants, where significance was considered at the 5%-level. Results are analyzed on a case-per-case basis in Sects. 3 and 4, while a more overall analysis is provided in Sect. 6.

3 Case 1: description and results

3.1 Description of the case

The first case relates to the bankruptcy of the (fictitious) real estate company Ghent Estate. The company is based in Ghent, which is a major city in Belgium. The objective of Ghent Estate is intermediating between sellers of luxury homes, who act as clients, and potential buyers. Participants were randomly presented one of three variants of this case. Common to the variants is the main role played by Jan Vandenbossche, one of the directors of Ghent Estate, who is held responsible for the bankruptcy by a majority of the clients due to certain inappropriate behavior on his behalf. Also common to the variants is the factual situation of Ghent Estate over time, in particular, that the company was established in 2008 and that its business was lucrative in the beginning. In all variants, Ghent Estate faced intense competition from Ghent Luxury Properties from 2019 onwards, in particular, due to the innovative approach of the latter company, eventually leading to the bankruptcy of Ghent Estate in 2021.

3.1.1 First variant

In the first variant, Jan Vandenbossche was convicted of fraud by the criminal court in 2019. It turned out that he had pressurized clients to pay higher commissions than initially agreed. He used the additional amount of money for personal expenses, especially to visit prostitutes, most of whom were minors. The subsequent liability claim by many clients resulted in reputational damage for Ghent Estate. It was estimated by Ghent Estate that the reputational damage resulted in a decrease of revenue by 20%.

Participants were asked to assess to what extent, in terms of the aforementioned Likert scale (Sect. 2.2), they agree that each of the following two facts are an explanation of the bankruptcy of Ghent Estate:

1. `The conviction of mister Vandenbossche and the associated reputational damage`

2. `The competition from Ghent Luxury Properties`

3.1.2 Second variant

The second variant also entails fraud committed by Jan Vandenbossche, but in this case the financial gain served an altruistic goal. In 2017 his wife was diagnosed with a terminal cancer, and it was only because his financial savings were virtually reduced to zero at a certain moment (due to the expensive cancer treatment to which he financially contributed), that Jan invented the plan to pressurize clients to pay higher commissions than contractually bound to. Furthermore, his fraud was restricted to clients who he suspected to enjoy a luxurious lifestyle. As in the first variant, many clients filed a liability claim after the fraud was discovered, resulting in reputational damage for Ghent Estate, and corresponding to—as estimated—a decrease of revenue by 20%.

The introduction of this variant was inspired by research results in the area of judicial decision-making, which indicates that irrelevant information, in particular litigant characteristics, may influence the decision (Wistrich et al. 2015). For example, in Liu and Li (2019) the authors presented participants with two variants of a fictitious legal case on contract law, where in the treatment condition the defendant was described as a person with bad moral character: she maintained an extramarital relationship with a corrupt government official. The experiment with real judges as participants revealed a statistically significant difference between the decisions of the two groups of participants, although the moral character was technically irrelevant to the case. In our case, the moral character also differs significantly between the two variants. The moral character is supposed to have limited influence on the occurrence of bankruptcy, since in both cases the objective effect of Jan's behaviour is a decrease of revenue by 20%. This variant thus allows to analyze whether the observed influence of moral character on legal decision-making is also present in a non-legal context, in particular a context related to economical affairs.

As in the first variant, participants were asked to assess to what extent they agree that each of the aforementioned two facts are an explanation of the bankruptcy of Ghent Estate. The first fact, however, was described in a slightly different way to take into account the change in moral character:

`Jan's plan to charge clients higher commissions than initially agreed, with as goal to finance his wife's cancer treatment, and the associated reputational damage`

3.1.3 Third variant

The third variant is a description of the same facts as the first variant, but (only) the order in which the information is presented differs. Whereas in the first variant the fraud by Jan Vandenbossche is outlined first, followed by the competition by Ghent Luxury Properties, the third variant starts

by describing the establishment of Ghent Luxury Properties and the associated competition.

This variant was introduced to test whether the perception of a situation depends on the order in which information is obtained and processed. This bears some similarity with the finding from psychological research that the temporality of events is important, in particular that people consider recent events more relevant or influential than more distal events (Miller and Gunasegaram 1990). Our variant allows us to investigate whether the same finding applies to the temporality of *the presented information about events*.

The facts that were presented to the participants as potential explanation of the bankruptcy of Ghent Estate, were described in the same way as in the first variant.

3.1.4 Explanations for bankruptcy in general

All participants were also asked to rate to what extent they agree that the implementation of given actions would be useful in obtaining explanations for the bankruptcy of *any* enterprise. The presented actions were the same for all variants, and are listed hereafter:

1. The development of AI software that produces an overview of the main financial details that relate to the bankruptcy of the given enterprise. The system is fully automatic, in the sense that there is no intervention by a human expert.

2. The development of AI software that identifies some other enterprises that did not go bankrupt. For each of the identified enterprises, the software indicates in what respect the financial details differ from those of the given bankrupt enterprise.

3. The development of AI software that returns as output the probability of bankruptcy. For its prediction, the involved algorithm relies on historic financial details from a large number of enterprises, where both bankrupt and non-bankrupt enterprises are taken into account.

4. When an enterprise encounters financial distress, the competent court summons a representative of the enterprise to obtain information on the state of affairs. If an enterprise eventually goes bankrupt, the previous statements of the representative are analyzed by the court.

The actions are presented to verify the research finding that many AI applications have limited take up, or are not appropriated at all, due to a lack of trust on behalf of their users (Linegang et al. 2006; Stubbs et al. 2007). The second and third action are introduced to investigate the more specific claim that users' trust in AI is increased if counterfactual explanations of the produced output are provided (Stepin et al. 2021; Chou et al. 2022). Such explanations address the need of people to understand why a certain event P happened *instead of* some other event Q, rather than knowing why event P happened (Miller 2019). The first and fourth actions present statements related to the presence or absence of human intervention. The description of the first action is intended to evaluate to what extent users accept fully automatic systems, while the last action relates to the other extreme of the absence of any computer assistance.

Below we will evaluate the interesting hypothesis that the responses to the degree of desirability of implementing AI software to increase explainability, are particularly found at the extremes of the Likert scale. This hypothesis is induced by recent research, where it is suggested that non-expert users tend to either over-trust or distrust AI software (Cohen et al. 2017; Larasati et al. 2021).

3.2 Results

3.2.1 Behavior of Jan Vandenbossche as an explanation of Ghent Estate's bankruptcy

As outlined above, participants were asked to give their opinion on the degree to which the behavior of Jan Vandenbossche contributed to the bankruptcy of Ghent Estate.

A bar chart of the responses is displayed in Fig. 2, and the mean value of the responses for each of the variants is shown in Table 3. It is obvious that, for all variants, there is a strong consensus among the respondents that the behavior of Jan Vandenbossche contributed to the bankruptcy of Ghent Estate. Furthermore, the figure and the table indicate no notable differences between the variants. This is confirmed by the application of the Kruskal–Wallis test, with p-values as shown in Table 4.

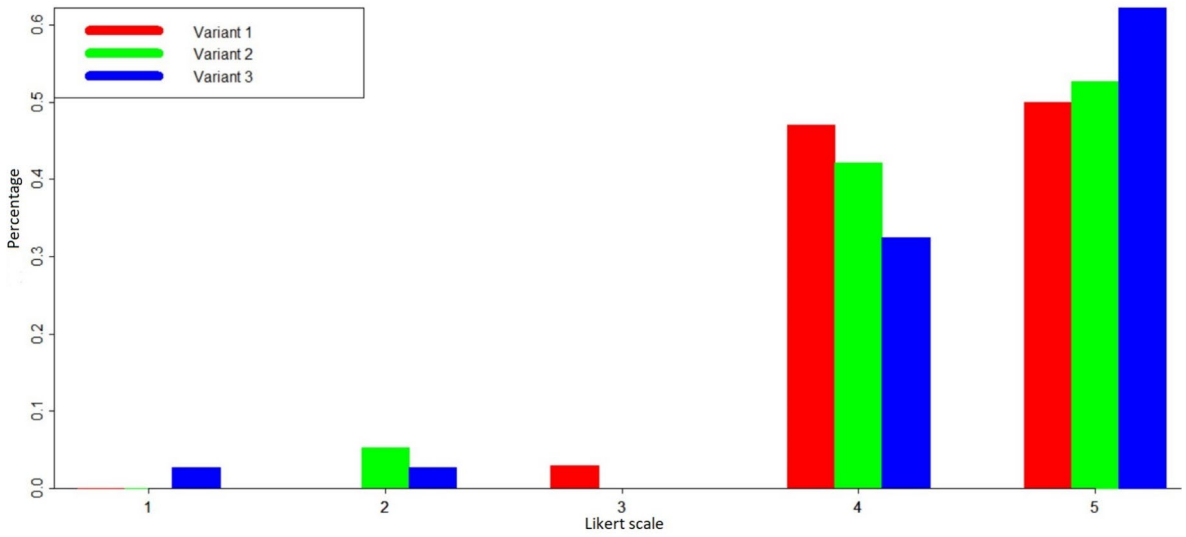


Fig. 2 Bar chart of the responses on the contribution of Jan Vandenbossche to the bankruptcy of Ghent Estate

Table 3 Mean value of the responses on the contribution of Jan Vandenbossche to the bankruptcy of Ghent Estate

Variant	Mean
1	4.47
2	4.42
3	4.48

Table 4 p-values of the Kruskal-Wallis test for the contribution of Jan Vandenbossche to the bankruptcy of Ghent Estate

	Variant 1	Variant 2	Variant 3
Variant 1		0.9184	0.3871
Variant 2			0.4509

3.2.2 Competition from Ghent Luxury Properties as the explanation of Ghent Estate's bankruptcy

Participants were also asked to give their opinion on the degree to which the competition from the recently established competitor Ghent Luxury Properties contributed to the bankruptcy of Ghent Estate. In all variants, Ghent Luxury Properties was established in 2019, and their innovative approach resulted in strong competition for Ghent Estate. In the first and the second variant, the competition from Ghent Luxury Properties was mentioned at the end of the case, i.e. the respondents were first presented facts about Ghent Estate, while the third variant started by presenting information about Ghent Luxury Properties.

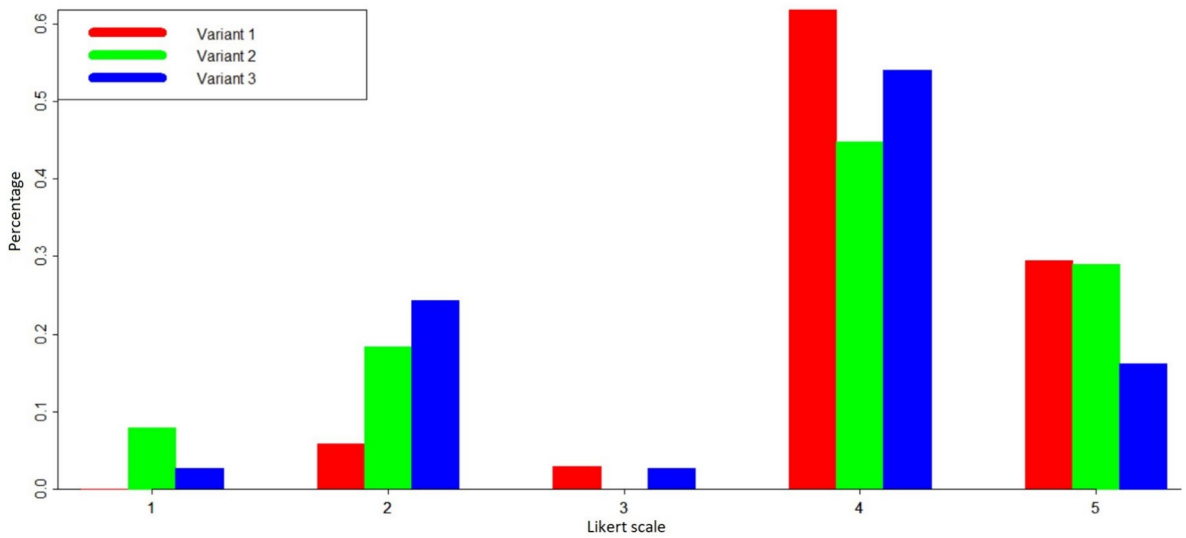


Fig. 3 Bar chart of the responses on the contribution of the competition from Ghent Luxury Properties to the bankruptcy of Ghent Estate

Table 5 Mean value of the responses on the contribution of the competition from Ghent Luxury Properties to the bankruptcy of Ghent Estate

Variant	Mean
1	4.14
2	3.68
3	3.56

Table 6 p-values of the Kruskal–Wallis test for the contribution of the competition from Ghent Luxury Properties to the bankruptcy of Ghent Estate

	Variant 1	Variant 2	Variant 3
Variant 1		0.2614	0.0248
Variant 2			0.4367

A bar chart of the responses is shown in Fig. 3, and the mean value of the responses for each of the variants is shown in Table 5. The p-values for the Kruskal-Wallis test are shown in Table 6.

The results show that there is a statistically significant difference between the first and the third variant. Since these variants differ *only* in the order in which information is presented, it follows that the perception of a situation might depend on the order in which information about this situation is processed.

The significant difference between the first and the third variant is not found between the second and the third variant, although the second variant presented the information about the case in the same order as the first variant. This observation might be explained by the fact that the first and the second variant are different in another respect, namely the behavior of Jan Vandenbossche. Whereas in the first variant Jan Vandenbossche was characterized as a rather unpleasant person, the second variant displayed him as a caring person. The difference in characterization thus apparently also affects the perception of a situation.

3.2.3 Explanations for bankruptcy in general

Respondents were also asked to evaluate certain actions that may, or may not, be helpful in increasing the explainability of bankruptcy in general. The proposed actions were the same for all variants. Bar charts of the responses are shown in Figs. 4, 5, 6 and 7.

Figure 4 indicates that most respondents mildly object the suggestion of the use of AI software that gives insight into the financial details that relate to the bankruptcy of a given enterprise, although there are also relatively many respondents who mildly agree that the presented suggestion might be useful. The figure is in line with the recent research result that was mentioned above, namely that non-expert users tend to either over-trust or distrust AI software (cf. Sect. 3.1.4). However, the figure shows that neither trust nor distrust are very pronounced (most responses are either 2 or 4 on the 5-point Likert scale).

Figure 5 shows that almost half of the respondents (mildly) agree that it would be useful to have AI software that compares a given bankrupt enterprise to non-bankrupt enterprises. This result can be considered as a confirmation of what has been observed in the literature: when it comes to explainability, users particularly like counterfactuals (cf. Sect. 3.1.4).

The bar chart in Fig. 6 is similar to the previous one. In particular, most respondents are in favor of the use of AI software that predicts the probability of bankruptcy. It might be hypothesized that the reason for this result is

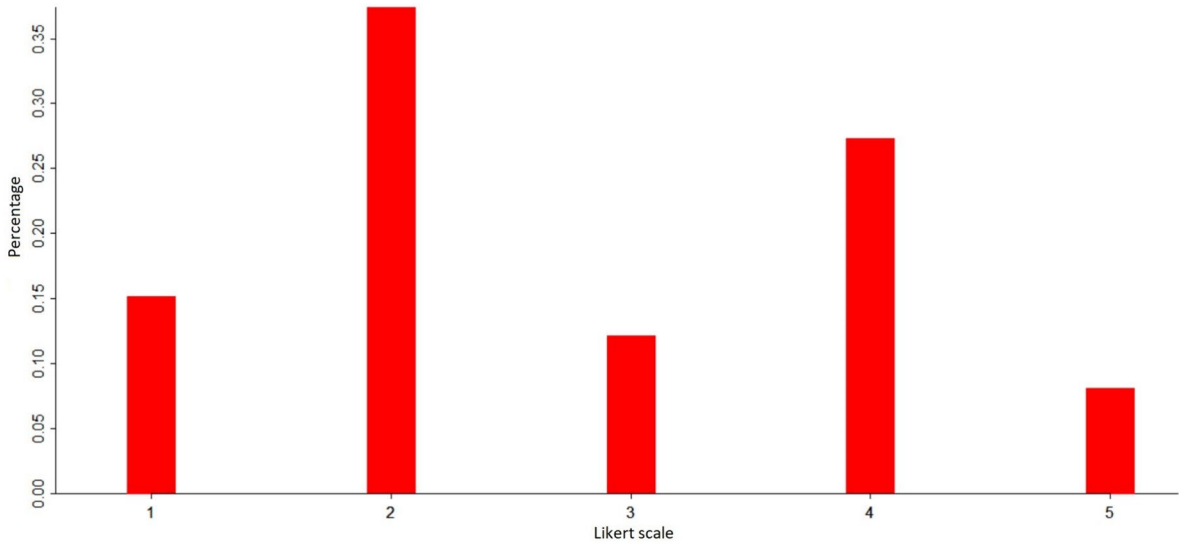


Fig. 4 Bar chart of the responses on the use of AI software to produce an overview of the financial details that relate to bankruptcy

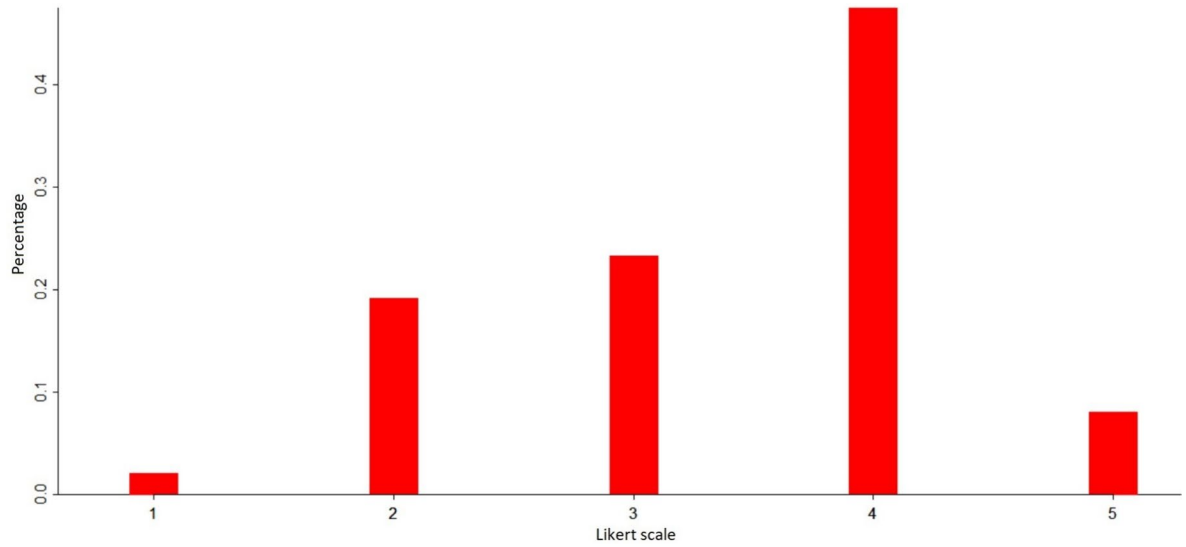


Fig. 5 Bar chart of the responses on the use of AI software to compare a given bankrupt enterprise to non-bankrupt enterprises

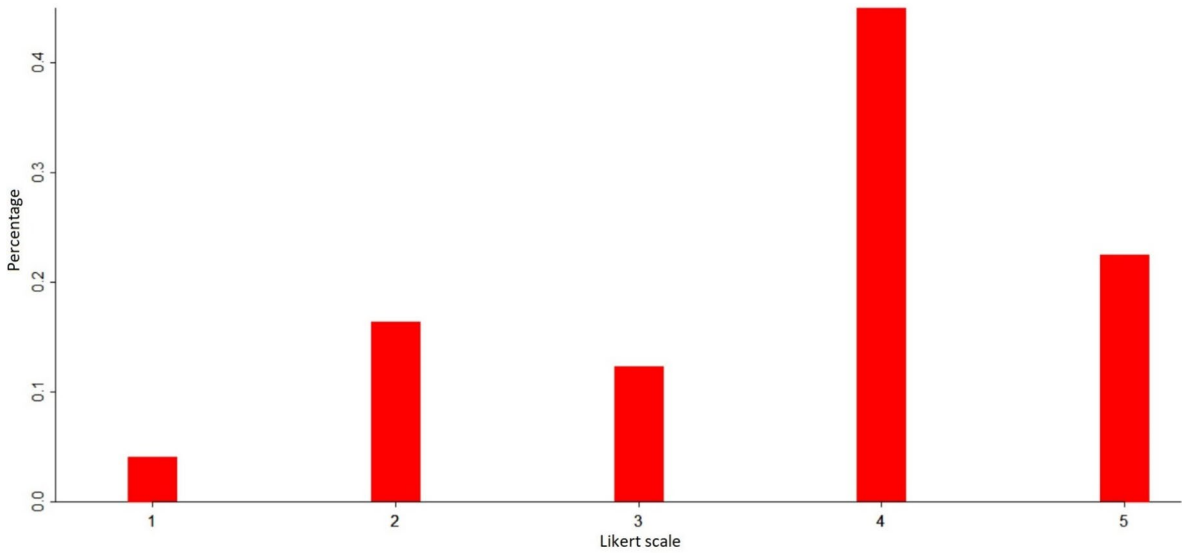


Fig. 6 Bar chart of the responses on the use of AI software to predict the probability of bankruptcy

also similar, namely that users appreciate explainability in terms of counterfactual situations. Interestingly, and in contrast to the previous suggested action, the counterfactuals in this case are implicit, in the sense that the corresponding financial details are not part of the produced output. Both bankrupt and non-bankrupt enterprises are taken into account in estimating the probability of bankruptcy of a given enterprise, but the financial details of these enterprises are not shown to the user. Thus it seems that users do not necessarily require to obtain specific knowledge on the counterfactuals, as long as they are ensured that counterfactuals are taken into account by the AI algorithm.

Finally, Fig. 7 clearly shows that most respondents appreciate the role of the commercial court. This indicates that human intervention is still of major importance for most users when it comes to obtaining explanations for certain events.

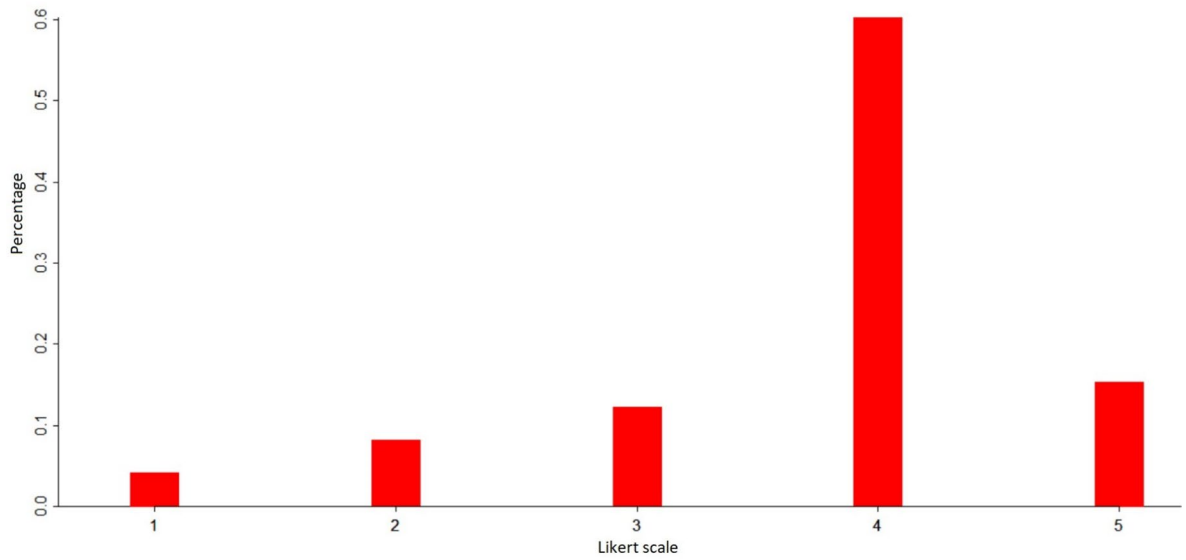


Fig. 7 Bar chart of the responses on the role of the commercial court

4 Case 2: description and results

4.1 Description of the case

In the second case, some more emotionally related aspects are introduced. Steven De Wachter is a young man who had a happy family life with his wife and daughter over a period of 10 years. He maintained a pretty stable life, notwithstanding his occasional abuse of drugs, a habit that was created during his time at the university.

However, since last year Steven seems like a different person. The unexpected and tragic death of his wife turned him into a sad person, displaying sudden episodes of unwarranted anger. Friends now see a person who is "no longer completely normal".

Six months ago Steven had another bad luck. He was scammed by someone he considered a great friend, the result being that the supposed friend, as well as virtually all Steven's money, have disappeared. One particular consequence of the financially unstable situation is that Steven has no other option than to reduce the frequency of trips with his little daughter.

Last month bad luck stroke again. The two variants of this case differ in the description of the unfortunate event that Steven experienced this time.

4.1.1 First variant

In the first variant, the unfortunate event of last month starts with Steven finding an interesting advertisement for a second-hand car at a very good price. Above all, the car seemed safer than his current one, and thus better suited as transport for his daughter. Incidentally, the car was offered by an old friend, so that he was very confident about the reliability of the car. Steven could not imagine that he would be scammed again by a friend.

Shortly after Steven left the seller with the newly bought car, it catches fire. Steven is barely able to stop the car and he jumps out of it. Understandably, Steven is furious, and he runs back to the seller, punching him several times in the face.

A few months later Steven is summoned to the correctional court for the injuries he caused to the seller. The judge turns out to be remarkably lenient with Steven.

4.1.2 Second variant

The second variant differs only in one subtle aspect with the first variant. Steven also finds an interesting advertisement for a second hand car at a very good price. Furthermore, the car is also, incidentally, offered by an old friend. However, the second variant does not mention that Steven's confidence in the reliability of the car is positively related to the fact that the seller is an old friend. Neither does the second variant contain the statement "Steven could not imagine that he would be scammed again by a friend."

The introduction of this variant allows us to analyze to what extent respondents are sensitive to suggestive but non-factual statements. Although neither variant contains facts from which it may be deduced that the seller scammed Steven, the statement "Steven could not imagine that he would be scammed again by a friend" from the first variant might arouse suspicion about the seller. In other words, it might induce the intuition or gut feeling in the respondents that the

seller is a scammer, which might affect their perspective on the case. This is known as the intuition bias (Sadler-Smith and Shefy 2004; Gosar and Solomon 2019).

4.1.3 Explanations for the mild sentence

In both variants, respondents were asked to imagine that Steven provided them the verdict and that he points to the surprisingly lenient sentence. Respondents were then invited to rate to what extent certain knowledge might be helpful in better understanding the judgment. It was stressed that respondents were supposed not to worry about how to actually obtain this knowledge; they just had to assume that they would receive the described information.

The additional information (i.e. in addition to the facts contained in the description of the case), to be rated by the respondents in terms of its use in increasing the explainability of the judgment, was as follows:

1. Knowledge of the sentence if Steven's wife would not have died.
2. Knowledge of the sentence if Steven would not have been scammed by the alleged friend, and thus would not have had financial difficulties.
3. Knowledge of the sentence in case Steven would have committed similar offenses in the past (i.e. in case of recidivism).
4. Knowledge of the articles of the Criminal Code on the basis of which Steven was convicted.
5. Knowledge of the case law of the criminal court regarding assault and battery, i.e. knowledge of verdicts for similar offenses in the past.

The first two pieces of information obviously relate to the use of counterfactuals in explainability (cf. Sect. 3.1.4). The given case allows to analyze the degree of appreciation of counterfactuals in the specific context of criminal law.

The other pieces of information relate to the currently prevailing view that explanations are ideally user-centered, i.e. different categories of users might require different types of explanations (Abdul et al. 2018; Ribera and Lapedriza 2019; Schoonderwoerd et al. 2021). The last three pieces of information are of particular relevance to persons with a background in law, so that it might be hypothesized that students in law are convinced that this information would increase explainability. Below it will also be analyzed whether the respondents have a preference for counterfactual explanations (the first two pieces of information) or for information that is directed towards experts (the last three pieces of information).

4.2 Results

Bar charts of the responses are shown in Figs. 8, 9, 10, 11 and 12. As described in Sect. 2.3, a twofold analysis is provided: one where all respondents are taken into account, and one where only responses from law students are taken into account. The mean values of the responses are displayed in Table 7 (rows correspond to each of the pieces of information described above).

The results indicate that for both variants each of the pieces of information is considered useful by most respondents since the mean values are at least 3.29. For each of the given pieces of information, the difference between the bar charts of both variants is rather minor. This was confirmed by the application of the Kruskal–Wallis test, which did not show any statistically significant difference.

Differences between the results where all respondents are taken into account versus the results where only law students are considered, are rather minor. Given the relatively small number of 'other' students (compared to the number of law students), it is not possible to establish whether these differences are due to different characteristics of both types of respondents or due to random variations.

Next, a comparison was made between the appreciation of each of the pieces of information *within* each of the variants. Table 8 contains the p-values for the differences between the pieces of information for the first variant, while Table 9 shows these values for the second variant. It is clear that when all respondents are taken into account, it holds that for the first variant it is significantly more appreciated to have knowledge of the sentence in case there would be recidivism, and to have knowledge of the relevant articles of the Criminal Code, than to have knowledge of the sentence in case Steven's wife would not have died. This observation does not hold when analysis is restricted to law students, but the statistical non-significance might simply be due to the smaller number of considered respondents. The result does, however, suggest that *all* types of respondents might appreciate explanations in terms of domain knowledge (at least in case the domain knowledge is not too technical or specialized).

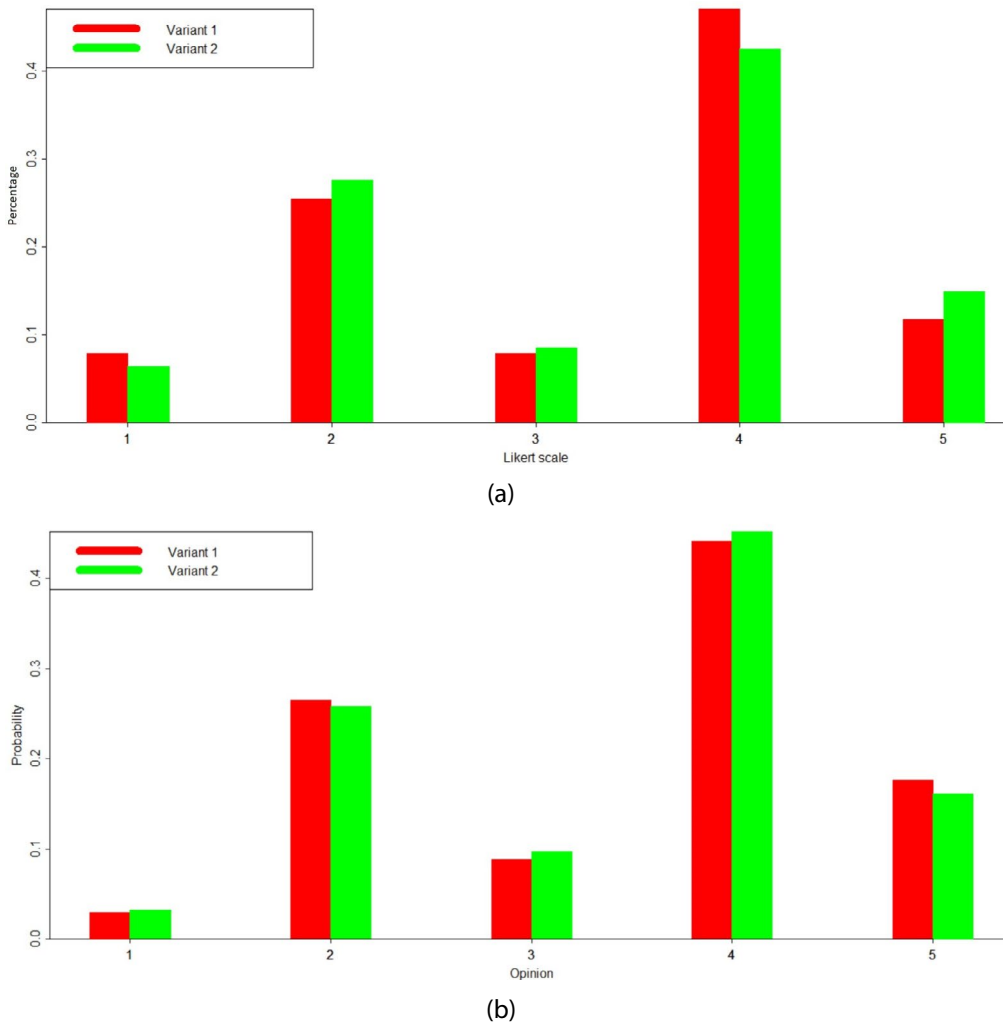


Fig. 8 Bar chart of the responses on the use of knowledge of the sentence in case Steven's wife would not have died, with **a** including all respondents, and **b** only law students

The same applies to the second variant, although for the second variant knowledge of relevant case law is also significantly more appreciated than knowledge of the sentence in case Steven's wife would not have died (and this is irrespective of the fact whether or not the results are restricted to law students). This result should not be interpreted as an indication that domain knowledge is more important than counterfactuals, since the aforementioned statistical differences are not found between the counterfactual of absence of scamming, on the one hand, and the domain information, on the other hand. However, the result at least suggests that users are critical about the counterfactuals that are presented to them, and that if some counterfactual is not considered very insightful, there is a strong preference for domain information instead.

5 Discussion of the research questions

Having described the results of the surveys, we return to the research questions from Sect. 2.1, and we also consider the significance of the results for real-world settings.

Current research on XAI is focused on how to provide explanations for automated decisions by AI systems, often by relying on other AI models. As a prototypical example, consider the popular model-agnostic interpretation method LIME (Ribeiro et al. 2016), which is mainly used in the area of image classification. LIME highlights the areas in the image that have been crucial for the prediction of a specific class. However, as stressed by Ghassemi et al. (2021), the important question for users trying to understand an individual decision is not

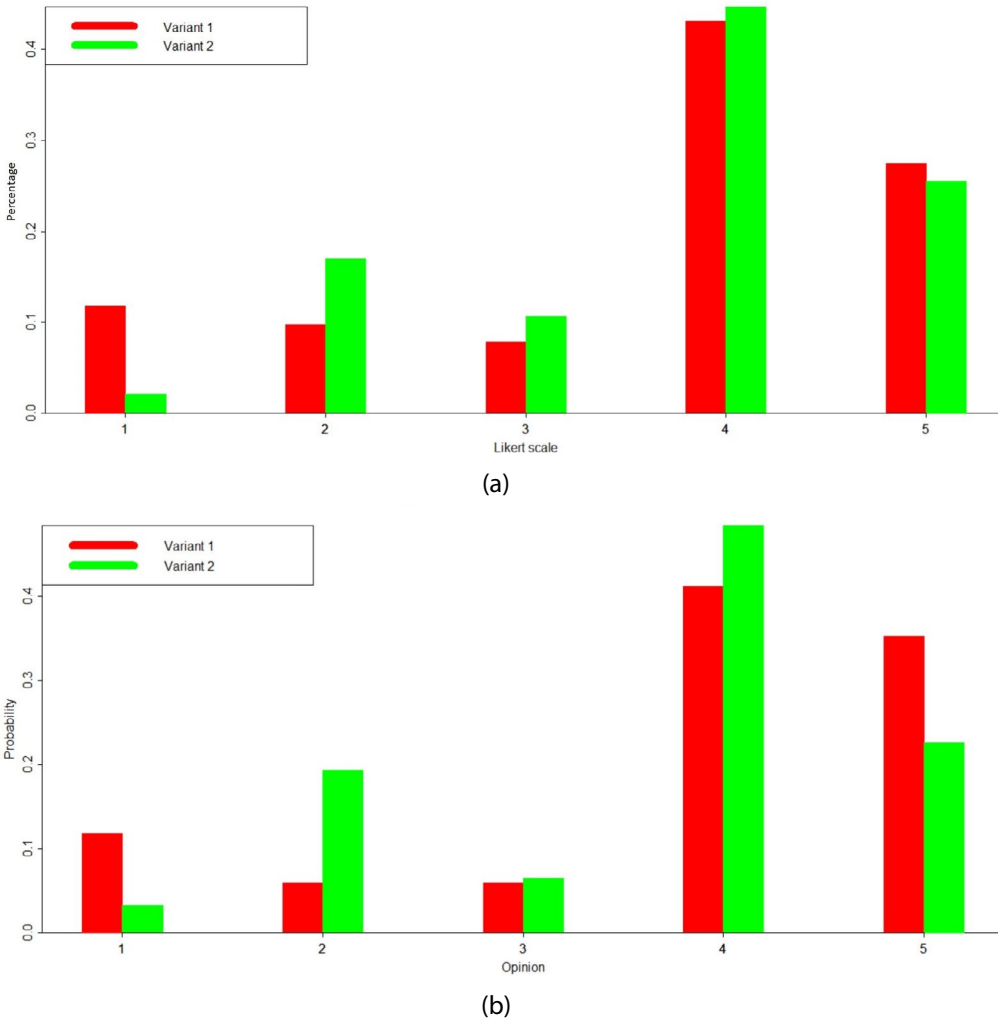


Fig. 9 Bar chart of the responses on the use of knowledge of the sentence in case Steven would not have been scammed, with **a** including all respondents, and **b** only law students

where the model was looking, but instead whether it was reasonable that the model was looking in that region. The same others argue, in the context of health care, that current explainability methods are unreliable to provide patient-level decision support. They conclude that presently the hope for human-comprehensible explanations for complex, black-box machine learning algorithms that can be used safely for bedside decision making, remains an open challenge.

As pertains to Q2, our findings align, to some extent, with the pessimistic view by Ghassemi and colleagues. Users, both expert and non-expert, seem to appreciate explanations in terms of domain knowledge, but current explanation methods are unable to provide explanations in terms of domain knowledge that have the same quality as the ones provided by human experts. Quality, in the foregoing sense, might be understood as a combination of comprehensibility, depth and relevance of the background knowledge, and communication skills to the user. Actually, the observation that a significant part of non-expert users distrust AI systems, might be due to the inability of explanation methods to generate understandable explanations that take background knowledge into account.

On the other hand, and addressing Q1, users also like counterfactual explanations, and generating such kind of explanations is a hot topic in current AI research. Taking the medical domain as a prototype for the importance of progress in XAI, it is reassuring that methods such as GANterfactual are being developed. The former method is able to generate counterfactual image explanations based on adversarial image-to-image translation techniques to enhance explanation in the medical context.

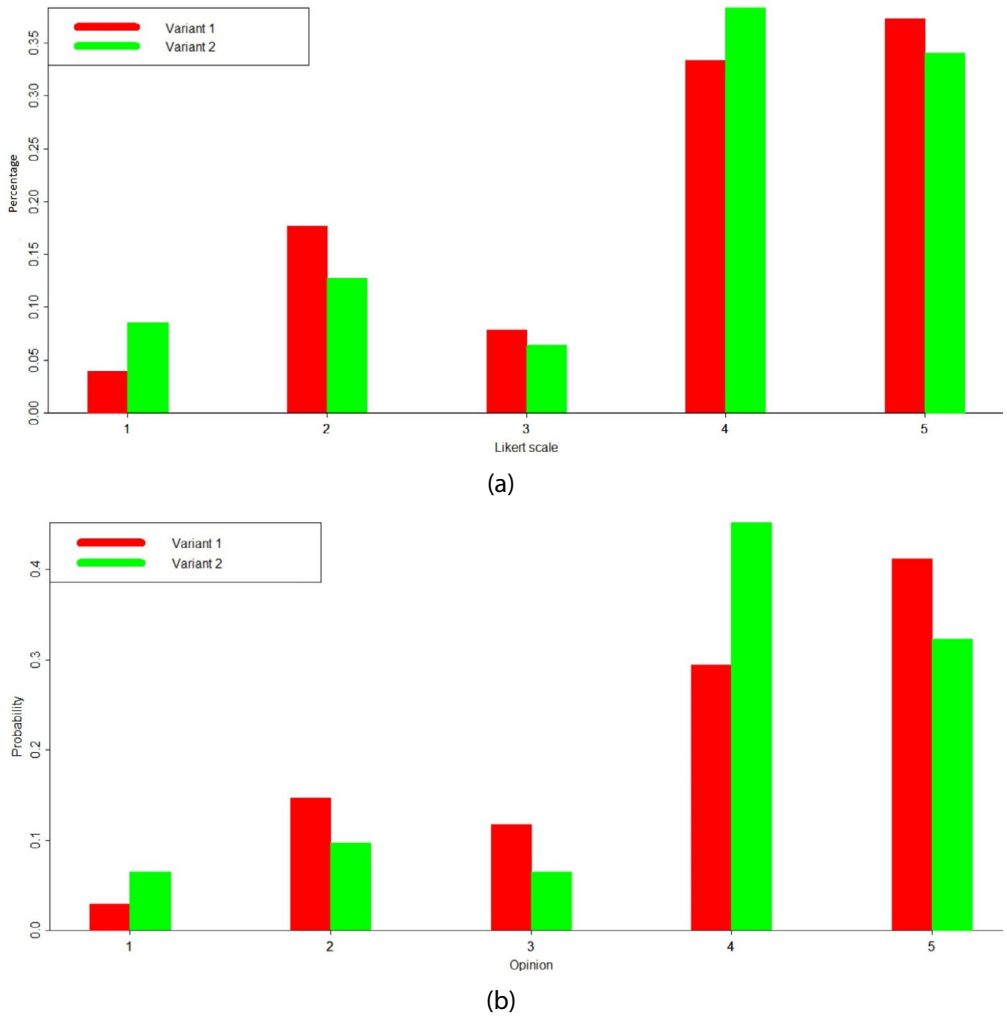


Fig. 10 Bar chart of the responses on the use of knowledge of the sentence in case there would be recidivism, with **a** including all respondents, and **b** only law students

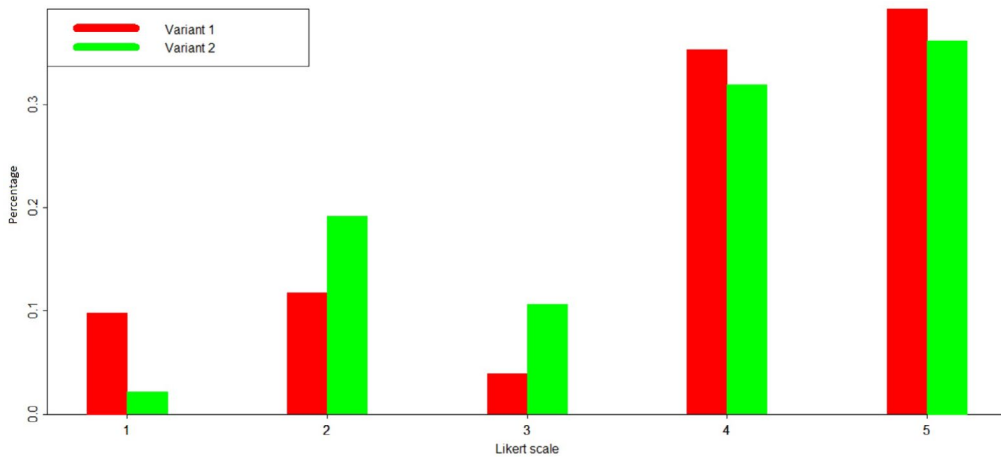
Considering Q3, the results of the first case do not show a significant influence of the moral behavior of the main person on the evaluation of the presented event. Indeed, the difference in behavior of Jan Vandenbossche between the variants of the first case did not result in a significant difference in the appreciation of the degree to which his behavior contributed to the bankruptcy. Since this result contradicts with what has been hypothesized in legal decision making (Liu and Li 2019), further research is needed to evaluate the impact of the moral presentation of a main character on the interpretation of an event.

In terms of Q4, users seem not to be prone to suggestive but non-factual statements as explanations of a given event. The variants of the second case were constructed to differ in the degree to which statements were suggestions rather than facts, but no statistical significant difference between the two variants of the second case was found.

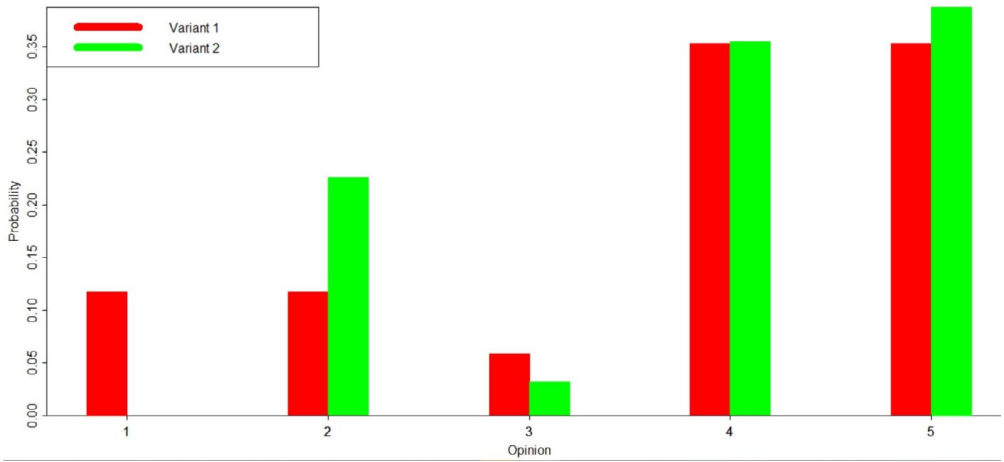
Considering Q5, our research suggests that the order in which explanations are presented, might effect the user's appreciation of AI-generated output. This observation is currently not taken into account in the current development of explanation methods, but it might be crucial in critical domains such as health care. For example, the patient's degree of hope or despair might be significantly affected by the order in which explanations of a certain diagnosis are presented.

6 Conclusion

This paper builds on recent research that integrates results from the social sciences with the field of XAI. The surveys that were presented to university students, mainly from the Faculty of Law, indicate some interesting insights. Some



(a)



(b)

Fig. 11 Bar chart of the responses on the use of knowledge of the articles of the Criminal Code on the basis of which Steven was convicted, with **a** including all respondents, and **b** only law students

results are rather preliminary and require further research, while other findings are confirmations of established knowledge. To the former category belong to the following hypotheses: 1. users are more concerned that computer models take counterfactuals into account in producing their output, than that these counterfactuals are actually presented to them as part of the output, 2. users recognize the use of AI in specialized domains, although this does not mean that human intervention is considered obsolete, 3. all users, experts users as well as non-expert users, appreciate explanations in terms of domain knowledge (at least if the domain knowledge is not too technical nor specialized), as well as explanations in terms of counterfactuals, 4. whether an explanation in terms of counterfactuals is considered more insightful than an explanation in terms of domain knowledge, might depend on both the details of the case and the extent to which the counterfactual is considered relevant, 5. the order in which facts are presented might be more important than the temporal order of the facts, and 6. users are not prone to suggestive but non-factual statements as explanations of a given event. Other findings confirm what researchers have found before, in particular that 1. counterfactuals play an essential role in understanding a given event, and 2. non-expert users tend to either over-trust or distrust AI. Finally, we did not find a significant influence of the moral behavior of the main person on the evaluation of the presented event.

These results are obviously relevant to the field of XAI. For example, the finding that users are prone to the order in which information is presented to them, irrespective of the temporal order of historical relevant facts, implies that computer scientists should carefully consider the format of the displayed output of any computer system.

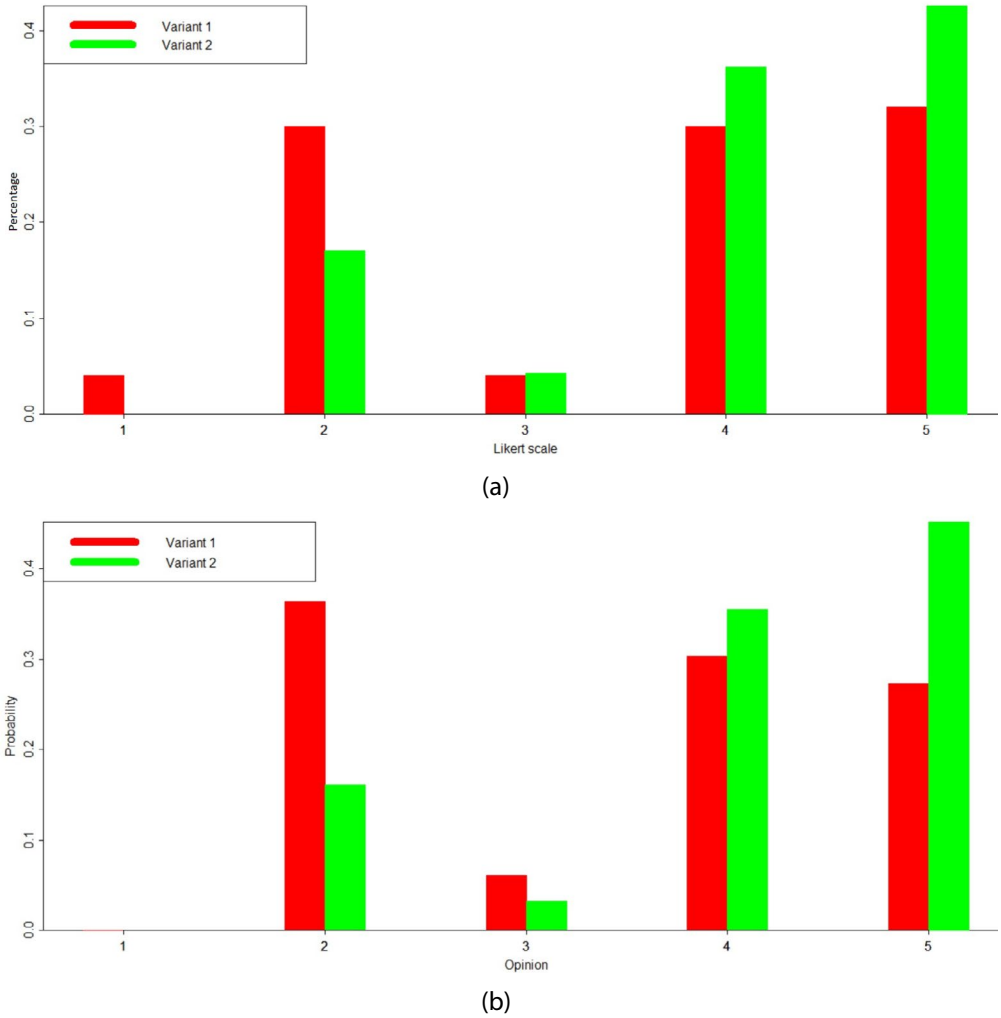


Fig. 12 Bar chart of the responses on the use of knowledge of relevant case law, with **a** including all respondents, and **b** only law students

Table 7 Mean value of the responses on the use of certain knowledge in understanding the conviction of Steven, with (a) including all respondents, and (b) only law students

	Variant 1	Variant 2
(a)		
1	3.29	3.32
2	3.65	3.74
3	3.82	3.77
4	3.82	3.81
5	3.47	4.04
(b)		
1	3.47	3.45
2	3.82	3.68
3	3.91	3.87
4	3.71	3.90
5	3.35	4.10

Table 8 p-values of the Kruskal–Wallis test for the first variant, with (a) including all respondents, and (b) only law students

	Wife	Scamming	Recidivism	Articles of law	Case law
(a)					
Wife		0.09	0.01	0.01	0.26
Scamming			0.43	0.32	0.74
Recidivism				0.87	0.27
Articles of law					0.25
(b)					
Wife		0.1302	0.08348	0.2636	0.9487
Scamming			0.8058	0.7957	0.1915
Recidivism				0.5886	0.1043
Articles of law					0.332

Table 9 p-values of the Kruskal-Wallis test for the second variant, with (a) including all respondents, and (b) only law students

	Wife	Scamming	Recidivism	Articles of law	Case law
(a)					
Wife		0.09	0.04	0.04	0.002
Scamming			0.64	0.62	0.12
Recidivism				0.93	0.30
Articles of law					0.35
(b)					
Wife		0.4045	0.1082	0.0851	0.01536
Scamming			0.4118	0.3169	0.08745
Recidivism				0.8108	0.3584
Articles of law					0.5174

Appendix

Below we outline the two cases in detail, together with the corresponding variants, as they were presented to the respondents. The questions corresponding to the cases are displayed in figures that follow the description of the related case.

The questions related to the implementation of given actions for explaining bankruptcy of any enterprise, as part of the first case, are not displayed below, as these questions were already fully described in Sect. 3.1.4.

Case 1

Variant 1

Real estate company Ghent Estate was founded in 2008 with the aim of selling luxurious houses in Ghent and surrounding municipalities. Ghent Estate's financial data were predominantly positive from the start.

In 2019, one of the directors, Jan Vandenbossche, was convicted by the criminal court after a complaint of fraud. The complaint arose in response to increasing suspicions by employees of Mr. Vandenbossche that he deviously put pressure on customers (who wanted to sell their house) to make them pay a higher commission than initially agreed. The part of the commission that was higher than the customer had originally agreed went to Mr. Vandenbossche. That extra income allowed Mr. Vandenbossche to lead a debauched life, especially as a customer of prostitutes, many of whom were underage and vulnerable girls. After the conviction of Mr. Vandenbossche, many former clients start filing liability claims against Ghent Estate, causing the company to suffer significant reputational damage. Ghent Estate estimates that, as a result, turnover has fallen by 20.

In 2019, a second setback occurred for Ghent Estate. That year, the competing company Ghent Luxury Properties was founded by a number of enthusiastic young entrepreneurs. Because of their innovative approach, they managed to convince many sellers of luxurious houses to choose their services.

Ghent Estate goes bankrupt in 2021 (Fig. 13).

Variant 2

Real estate company Ghent Estate was founded in 2008 with the aim of selling luxurious houses in Ghent and surrounding municipalities. Ghent Estate's financial data were predominantly positive from the start.

In 2017, the wife of one of the directors, namely director Jan Vandenbossche, was diagnosed with terminal cancer. Jan is a dutiful husband and therefore makes excessive efforts, both financially and practically, to support his wife. However, Jan encounters financial problems. Two years earlier, his previous wife left him, and she managed—thanks to her superb lawyer—to legally oblige him to pay a very high alimony for their mutual children. Moreover, his wife's cancer treatment costs a lot

Indicate to what extent you agree or disagree that the facts presented below provide an explanation for the bankruptcy of Ghent Estate. **Select one box in each row.**

	Strongly disagree	Slightly disagree	No opinion	Slightly agree	Strongly agree
1. The conviction of mister Vandenbossche and the associated reputational damage	☐	☐	☐	☐	☐
2. The competition from Ghent Luxury Properties	☐	☐	☐	☐	☐

Fig. 13 Questions presented to respondents for case 1, variant 1

Indicate to what extent you agree or disagree that the facts below provide an explanation for the bankruptcy of Ghent Estate. **Select one box in each row.**

	Strongly disagree	Slightly disagree	No opinion	Slightly agree	Strongly agree
1. Jan's plan to charge clients higher commissions than they had initially agreed to, in order to finance his wife's cancer treatment, and the associated reputational damage	☐	☐	☐	☐	☐
2. The competition from Ghent Luxury Properties	☐	☐	☐	☐	☐

Fig. 14 Questions presented to respondents for case 1, variant 2

Indicate to what extent you agree or disagree that the facts presented below provide an explanation for the bankruptcy of Ghent Estate. **Select one box in each row.**

	Strongly disagree	Slightly disagree	No opinion	Slightly agree	Strongly agree
1. The conviction of mister Vandenbossche and the associated reputational damage	☐	☐	☐	☐	☐
2. The competition from Ghent Luxury Properties	☐	☐	☐	☐	☐

Fig. 15 Questions presented to respondents for case 1, variant 3

of money. At a certain moment, his savings are virtually reduced to zero. Desperate, Jan invents a plan where he charges higher commissions to customers (who want to sell their house) than originally agreed. He used the part of the commission that was higher than originally agreed to continue to pay his wife's cancer treatment. However, because his plan conflicts with his strong sense of justice, Jan only applies it to customers who he suspects to have a very luxurious lifestyle. However, in 2019 his plan is discovered, with the result that many former customers file a liability claim, causing Ghent Estate to suffer significant reputational damage. Ghent Estate estimates that, as a result, turnover has fallen by 20

In 2019, a second setback occurred for Ghent Estate. That year, the competing company Ghent Luxury Properties was founded by a number of enthusiastic young entrepreneurs. Because of their innovative approach, they manage to convince many sellers of luxurious houses to choose their services.

Ghent Estate goes bankrupt in 2021 (Fig. 14).

Variant 3

Real estate company Ghent Estate was founded in 2008 with the aim of selling luxurious houses in Ghent and surrounding municipalities. Ghent Estate's financial data were predominantly positive from the start.

However, in 2019, the competing company Ghent Luxury Properties was founded by a number of enthusiastic young entrepreneurs. Because of their innovative approach, they manage to convince many sellers of luxurious houses to choose their services.

In 2019, a second setback occurred for Ghent Estate. In that year, one of the directors, namely Jan Vandenbossche, was convicted by the criminal court after a complaint of fraud. The complaint arose in response to increasing suspicions by employees of Mr. Vandenbossche that he deviously put pressure on customers (who wanted to sell their house) to make them pay a higher commission than initially agreed. The part of the commission that was higher than the customer had originally agreed went to Mr. Vandenbossche.

Steven gives you a copy of the verdict, and he points out the surprisingly mild punishment. Please indicate to what extent knowledge of each of the following pieces of information would contribute to a better understanding of the judge's decision. This concerns additional information, i.e. additional to the case outlined. You should not wonder whether you could actually obtain that information; you therefore assess the usefulness of the additional data on the assumption that you *would* obtain that data. **Select one bow in each row.**

	Very useless	Slightly useless	No opinion	Slightly useful	Very useful
1. Knowledge of the sentence the judge would have imposed if Steven's wife had not died.	☐	☐	☐	☐	☐
2. Knowledge of the sentence that the judge would have imposed if Steven had not been defrauded by the alleged friend and therefore would not have had financial problems.	☐	☐	☐	☐	☐
3. Knowledge of the sentence that the judge would have imposed if Steven had already committed similar offenses in the past (recidivism)	☐	☐	☐	☐	☐
4. Knowledge of the articles of the Criminal Code on the basis of which Steven was convicted	☐	☐	☐	☐	☐
5. Knowledge of the case law of the criminial court regarding assault and battery, which means that you would receive an overview of what sentences judges have given in the past for similar offenses	☐	☐	☐	☐	☐

Fig. 16 Questions presented to respondents for case 2, variant 1

That extra income allowed Mr. Vandenbossche to lead a debauched life, especially as a customer of prostitutes, many of whom were underage and vulnerable girls. After the conviction of Mr. Vandenbossche, many former clients start filing liability claims against Ghent Estate, causing the company to suffer significant reputational damage. Ghent Estate estimates that, as a result, turnover has fallen by 20.

Ghent Estate goes bankrupt in 2021 (Fig. 15).

Case 2

Variant 1

Steven De Wachter is a young man who had a happy family life with his wife and daughter for ten years. He managed to lead a stable life, despite his sporadic use of drugs, a bad habit that he can barely get under control since he came into contact with fellow students during his university studies who sometimes offered him a joint.

However, since last year, Steven is no longer the same person. The sudden and tragic death of his wife had the consequence that he is rarely seen laughing and having fun. On the contrary, friends and acquaintances have noticed that he has had sudden fits of anger since then. They describe him as someone who is 'not completely normal anymore' since the death of his wife.

Six months ago, disaster struck again. Steven was scammed by someone he considered a close friend. Both the alleged friend and his savings have vanished since the incident. Now he encounters financial problems, and he even had to cut back on fun outings with his daughter.

Last month Steven saw an interesting advertisement where a nice second-hand car was offered for a very low price. The car seemed much safer than his current one, and, therefore, better suited to transport his daughter. Coincidentally, the car was offered for sale by an old acquaintance, so Steven was very confident about it. He couldn't imagine being deceived again by an acquaintance.

A week after paying the agreed amount, Steven goes to the acquaintance to pick up the car. Steven has only just left when the car catches fire. Fortunately, Steven is still able to brake and jump out of the car. Furiously, he walks back to the old acquaintance and punches him several times in the face.

A few months later, Steven is summoned to the criminal court for assault and battery. The judge gives him a surprisingly lenient sentence (Fig. 16).

Steven gives you a copy of the verdict, and he points out the surprisingly mild punishment. Please indicate to what extent knowledge of each of the following pieces of information would contribute to a better understanding of the judge's decision. This concerns additional information, i.e. additional to the case outlined. You should not wonder whether you could actually obtain that information; you therefore assess the usefulness of the additional data on the assumption that you *would* obtain that data. **Select one bow in each row.**

	Very useless	Slightly useless	No opinion	Slightly useful	Very useful
1. Knowledge of the sentence the judge would have imposed if Steven's wife had not died.	☐	☐	☐	☐	☐
2. Knowledge of the sentence that the judge would have imposed if Steven had not been defrauded by the alleged friend and therefore would not have had financial problems.	☐	☐	☐	☐	☐
3. Knowledge of the sentence that the judge would have imposed if Steven had already committed similar offenses in the past (recidivism)	☐	☐	☐	☐	☐
4. Knowledge of the articles of the Criminal Code on the basis of which Steven was convicted	☐	☐	☐	☐	☐
5. Knowledge of the case law of the criminial court regarding assault and battery, which means that you would receive an overview of what sentences judges have given in the past for similar offenses	☐	☐	☐	☐	☐

Fig. 17 Questions presented to respondents for case 2, variant 1

Variant 2

Steven De Wachter is a young man who had a happy family life with his wife and daughter for ten years. He managed to lead a stable life, despite his sporadic use of drugs, a bad habit that he can barely get under control since he came into contact with fellow students during his university studies who sometimes offered him a joint.

However, since last year, Steven is no longer the same person. The sudden and tragic death of his wife had the consequence that he is rarely seen laughing and having fun. On the contrary, friends and acquaintances have noticed that he has had sudden fits of anger since then. They describe him as someone who is 'not completely normal anymore' since the death of his wife.

Six months ago, disaster struck again. Steven was scammed by someone he considered a close friend. Both the alleged friend and his savings have vanished since the incident. Now he encounters financial problems, and he even had to cut back on fun outings with his daughter.

Last month Steven saw an interesting advertisement where a nice second-hand car was offered for a very low price. Steven saw a great opportunity to finally reduce the financial problems for him and his daughter: he would buy the advertised car and sell his current car, thus making a profit. Coincidentally, the car was offered for sale by an old acquaintance.

A week after paying the agreed amount, Steven goes to the acquaintance to pick up the car. Steven has only just left when the car catches fire. Fortunately, Steven is still able to brake and jump out of the car. Furiously, he walks back to the old acquaintance and punches him several times in the face.

A few months later, Steven is summoned to the criminal court for assault and battery. The judge gives him a surprisingly lenient sentence (Fig. 17).

Acknowledgements This work was supported by the Research Foundation - Flanders (Grant number G006421N).

Funding This work was funded by the Research Foundation - Flanders (Grant number G006421N)

Data availability The data is available upon request.

References

Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M (2018) Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp 1–18

Adadi A, Berrada M (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160

Chang L (1994) A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity. Appl Psychol Meas 18:205–215

Chou Y, Moreira C, Bruza P, Ouyang C, Jorge J (2022) Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. Information Fusion 81:59–83

Cohen MR, Smetzer JL, Miller T (2017) ISMP Medication Error Report Analysis: Understanding Human Over-reliance on Technology It's Exelan, Not Exelon Crash Cart Drug Mix-up Risk with Entering a "Test Order". Hospital Pharmacy 52:7–12 (2019)

Ghassemi M, Oakden-Rayner L, Beam A (2021) The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. The Lancet Digital Health 3:e745–e750

Girotto V, Legrenzi P, Rizzo A (1991) Event Controllability in Counterfactual Thinking. Acta Physiol (Oxf) 78:111–133

Gosar A, Solomon R (2019) Literature review on the Role of Intuition in Decision Making Process. World Journal of Research and Review 9:4–8

Joshi A, Kale S, Chandel S, Pal DK (2015) Likert Scale: Explored and Explained. British Journal of Applied Science & Technology 7:396–403

Kruskal WH, Wallis WA (1952) Use of Ranks in One-criterion Variance Analysis. Journal of the American Statistical Association, 47:583–621. Chang, L. (1994). A Psychometric Evaluation of 4-Point and 6-Point Likert-Type Scales in Relation to Reliability and Validity. Applied Psychological Measurement, 18:205–215. Joshi, A., Kale, S., Chandel, S., Pal, D.K. (2015). Likert Scale: Explored and Explained. British Journal of Applied Science & Technology 7:396–403

Larasati R, De Liddo A, Motta E (2021) AI Healthcare System Interface: Explanation Design for Non-Expert User Trust. In: Glowacka Dorota, Krishnamurthy Vinayak (eds) ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops. CEUR Workshop Proceedings, 2903

Linegang MP, Stoner HA, Patterson MJ, Seppelt BD, Hoffman JD, Crittendon ZB, Lee JD (2006) Human-automation Collaboration in Dynamic Mission Planning: a Challenge Requiring an Ecological Approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50:2482–2486

Liu JZ, Li X (2019) Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges. J Empir Leg Stud 16:630–670

Mertes S, Huber T, Weitz K, Heimerl A, André E (2022) GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. Frontiers in Artificial Intelligence 5

Miller T (2019) Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif Intell 267:1–38

Miller DT, Gunasegaram S (1990) Temporal Order and the Perceived Mutability of Events: Implications for Blame Assignment. J Pers Soc Psychol 59:1111–1118

Ribeiro M, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1135–1144

Ribera M, Lapedriza A (2019) Can We Do Better Explanations? A Proposal of User-Centered Explainable AI. In: Joint Proceedings of the ACM IUI 2019 Workshops

Rudin C, Radin J (2019) Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review 1

Sachan S, Yang J, Xu D, Benavides D, Li Y (2020) An explainable AI Decision-support-system to Automate Loan Underwriting 144

Sadler-Smith E, Shefy E (2004) The Intuitive Executive: Understanding and Applying 'Gut Feel' in Decision-Making. Decision-Making and Firm Success 18:76–91

Schoonderwoerd T, Jorritsma W, Neerincx M, van den Bosch K (2021) Human-centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. International Journal of Human-Computer Studies 154

Stepin I, Alonso J, Catala A, Pereira-Fariña (2021) A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. IEEE Access 9:11974–12001

Stubbs K, Hinds P, Wettergreen D (2007) Autonomy and Common Ground in Human-robot Interaction: a Field Study. IEEE Intell Syst 22:42–50

Thompson N, Spanuth S (2018) The Decline of Computers As a General Purpose Technology: Why Deep Learning and the End of Moore's Law are Fragmenting Computing. Information Sciences & Economics eJournal

Trabasso T, Bartolone J (2003) Story Understanding and Counterfactual Reasoning. J Exp Psychol Learn Mem Cogn 29:904–923

Wang D, Yang Q, Abdul A, Lim B (2019) Designing Theory-Driven User-Centric Explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems

Wistrich A, Rachlinski J, Guthrie C (2015) Heart versus Head: Do Judges Follow the Law or Follow Their Feelings. Texas Law Review 93:855–923