Artificial Intelligence for Engineering Design, Analysis and Manufacturing

cambridge.org/aie

Procedure for assessing the quality of explanations in failure analysis

Kristian González Barman

Centre for Logic and Philosophy of Science, Department of Philosophy and Moral Sciences, Ghent University, Blandijnberg 2, 9000 Ghent, Belgium

Abstract

This paper outlines a procedure for assessing the quality of failure explanations in engineering failure analysis. The procedure structures the information contained in explanations such that it enables to find weak points, to compare competing explanations, and to provide redesign recommendations. These features make the procedure a good asset for critical reflection on some areas of the engineering practice of failure analysis and redesign. The procedure structures relevant information contained in an explanation by means of structural equations so as to make the relations between key elements more salient. Once structured, the information is examined on its potential to track counterfactual dependencies by offering answers to *relevant* what-if-things-had-been-different questions. This criterion for explanatory goodness derives from the philosophy of science literature on scientific explanation. The procedure is illustrated by applying it to two case studies, one on Failure Analysis in Mechanical Engineering (a broken vehicle shaft) and one on Failure Analysis in Civil Engineering (a collapse in a convention center). The procedure offers failure analysts a practical tool for critical reflection on some areas of their practice while offering a deeper understanding of the workings of failure analysis (framing it as an explanatory practice). It, therefore, allows to improve certain aspects of the explanatory practices of failure analysis and redesign, but it also offers a theoretical perspective that can clarify important features of these practices. Given the programmatic nature of the procedure and its object (assessing and refining explanations), it extends work on the domain of computational argumentation.

Research Article

Cite this article: González Barman K (2022). Procedure for assessing the quality of explanations in failure analysis. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **36**, e26, 1–14. https://doi.org/10.1017/S0890060422000099

Received: 15 July 2021
Revised: 13 January 2022
Accepted: 2 April 2022

Key words:
Computational argumentation; epistemology; failure analysis; redesign; structural equation modeling

Author for correspondence:
Kristian González Barman, E-mail: kristiancampbell.gonzalezbarman@ugent.be

Introduction

Failure analysis is one of the main pillars of redesign. It allows to learn from mistakes and to improve current designs. One of the foundations of many failure analyses consists in their underlying accident causation models (ACMs). ACMs are structures that encode assumptions on the causality of accidents.[1] These assumptions lead to different accident investigation methods[2] (where each ACM can have different investigation methods), although not every investigation method has an associated ACM (Katsakiori *et al.*, 2009). There are also different systems and tools to represent (and sometimes to validate) the data gathered by these accident investigation methods.

The first ACMs were rather simple (a single causal chain leading from a single cause to a failure in a domino-like fashion) but have become increasingly more sophisticated, integrating multiple causes, and representing complex situations. Some authors consider, given the level of complexity of current technologies (and in some instances, socio-technological aspects), that event-based models should be substituted for models that consider engineering systems. A prime example is Nancy Leveson, for whom "[t]he cause of an accident, instead of being understood in terms of a series of events, is viewed as the result of a lack of constraints imposed on the system design and on operations, that is, by inadequate enforcement of constraints on behavior at each level of a socio-technical system" (2004, p. 251).

ACMs influence how engineers gather data and constrain the space of possible explanations given to a failure. Most methods deliver a set of causal factors (sometimes weighed and/or connected by a causal tree-like structure) that can make sense of a failure. Using these factors, it becomes possible to create an *explanation* of the failure, that is, to select relevant factors and connect them appropriately in order to answer an explanatory question. This means that there

[1] Several classifications have been suggested. Laflamme (1990) divided accident causation models into decisional, sequential, energetic and sequential, and organizational models; Lehto and Salvendy (1991) into general models of the accident process, models of human error and unsafe behavior, and models of human injury mechanics; Hollnagel (2002) into sequential accident models, epidemiological accident models, and systemic accident models.

[2] Some examples of investigation methods include: Hendrick and Benner (1987), Kleer and Williams (1987), Goel and Chandrasekaran (1989), Josephson and Josephson (1994), Chantler *et al.* (1995), Reiter (1998), Abdelhamid and Everett (2000), Dennies (2002), Affonso (2006), Andersen and Fagerhaug (2006), Bell *et al.* (2007), Xing and Amari (2008), Bhaumik (2009), Li *et al.* (2009), and Jensen *et al.* (2014).

can be different explanations for the same failure that utilize the same causal model. Similarly, given a certain research question, some explanations will be better than others, even if all candidate-explanations are validated by evidence. An explanation may be valid, in the sense of being correct (i.e., being supported by evidence), but nonetheless it may not offer the right kind of information to answer questions that might be deemed important. For example, an explanation aiming at the adjudication of legal responsibility will focus on different causal factors than an explanation attempting to replace a failed component, hence it will be likely that the former explanation will not provide adequate information for the aim of the latter and vice versa.

The variety of ACMs and methodologies has naturally led to discussions on the appropriateness of each method and has delivered different evaluation metrics to approach the issue (cf. Wagenaar and van der Schrier, 1997; Sklet, 2004; Katsakiori *et al.*, 2009; Saleh *et al.*, 2010). Most of these evaluations tackle the issue of appropriate methodologies and the validation thereof, but little has been said of the explanatory component.[3] For instance, Stern and Luger (1997) talk about (abductively) creating explanations using schemas, but they do not then consider the adequacy of the resulting explanations in terms of how well they satisfy their aims (or whether they could satisfy different aims). While it is important to have valid models and techniques, it is also important to consider the adequacy of the explanations that stem from said models. This point is sometimes trivial, since the validation of a model is normally motivated by a certain research question and is therefore built in such a way that validating it guarantees an adequate answer to said question; but this need not always be the case. In many instances, the explanation given for a specific failure might be imported into a different context (whence it originated); it could also be extended beyond its initial intended scope; or it might be used within a different discipline. Furthermore, answering an explanatory question is usually not a binary issue, but rather a matter of degree, where it is often possible to improve an already acceptable explanation.

The primary aim of this paper is to provide a procedure by which to evaluate the adequacy of failure explanations by measuring how well these explanations achieve their goal (i.e., answer certain explanatory questions). In this sense, the objective of the procedure is not to validate a certain model or explanation, but to assess its adequacy with regards to a specific problem (where "adequacy" refers to how the explanations fit what the target audience expects and needs). This aim is addressed through a five-step procedure. Reaching this chief aim directly leads to a *secondary aim*: showing how thinking of failure analysis as an explanatory practice offers a richer understanding of the field. Failure analysis is often said to be concerned with finding the causes of failure (see quotes in the section "Failure analysis as an explanatory practice"), and while this is partly true, a richer perspective that considers failure analysis as building explanations of failure might be more useful. The expectation motivating this paper is that, aside from improving understanding of the workings of failure analysis, the development of an evaluation tool will be an asset in comparing and refining the adequacy of certain types of failure explanations. This means that the paper has both a theoretical and a practical scope of application.

The first two steps of the procedure are instrumental for the rest, where the first step emphasizes the content of the explanation and the second emphasizes the ideal result we would expect from an optimal explanation. The third and fourth steps confront the ideal (optimal) results (i.e., ones with the ability to provide answers to the set of relevant w-questions) with the actual ones; more specifically, step 3 weighs the actual result against the ideal one in order to assess the performance of the actual explanation, and step 4 compares the results of the actual explanation against other possible explanations. Finally, the fifth step allows to validate (or not) the current explanation in light of the previous steps and gives redesign recommendations if pertinent.

The instruments used to develop this procedure are concepts and ideas from the literature on structural equation modeling (SEM) and from the literature on scientific explanation within the philosophy of science. In order to study and compare explanations, step 1 (*transcription*) structures the information they contain using SEM. SEM can be profitably used to abstract from details and capture fundamental features of explanations. In the case of the proposed procedure, it can be used to convert large portions of descriptive text into a few equations representing how variables (representing features of engineered mechanisms systems and their environment) relate to one another.

In step 2 (*identification*), the aim of the explanation is identified. We can then proceed to examine how well the structural equations produced in step 1 fulfill this aim. Such an examination is carried out in step 3 (*exploration and corroboration*) using a principle derived from the philosophical literature on scientific explanation (Woodward, 2003; Ylikoski and Kuorikoski, 2010): an adequate explanation should enable the tracking of systematic patterns of counterfactual dependence by answering what-if-things-had-been different questions (w-questions). The quality of an (adequate) explanation improves with its ability to answer more *relevant* w-questions. w-questions ask how a situation would had been different had there been different initial conditions. The "relevancy" of a w-question refers to whether its answer helps achieving the aim of the explanation or not. For instance, when a plane crashes during very bad weather conditions and an explanation aims at providing safety recommendations for future flights, it seems relevant to assess whether the plane would have also crashed in the absence of bad weather conditions (or to what degree the weather played a role), while it is irrelevant to assess whether the plane would have also crashed had it had a different color.

Step 4 (*comparison*) uses the same tools as step 3 to compare different explanations (i.e., contrasting sets of answerable relevant w-questions). Moreover, it allows to compare a single explanation with modified versions of itself, thereby promoting optimization. Step 5 (*validation and recommendation*) uses the results from steps 3 and 4 in order to validate an explanation: an explanation is validated if it is of a good quality (it answers the set of w-questions relevant to the aim) and it is optimized. If the aim of the explanation is redesign, it is possible to use the structural equations to provide useful redesign information (in terms of answers to interesting w-questions that consider the outcome of changes in design parameters).

The paper focuses on simple cases of explanations that are mostly event-based and do not involve layers of human interaction in order to properly illustrate the workings of the procedure in two separate instances. This does not mean that SEM cannot

[3] Also, in the philosophy literature, where the last few decades have revealed an increasing interest in the explanation practices of engineering sciences (Gabbay *et al.*, 2009; Barman and van Eck, 2021), most analyses focus on the structure of explanations (e.g., de Ridder, 2006; Boon, 2008) rather than on providing an account of, or criteria for, the quality or adequacy (or lack thereof) of such explanations.

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

handle more complex systems (it is quite the contrary, given the ability of structural equations to deal with complex relations between variables and latent variables). SEM allows to extract counterfactuals while remaining neutral with respect to which accident causation model underlies the explanation. SEM is, therefore, a useful tool to capture complex situations in a rigorous manner, especially when certain details are not fleshed out (since it can incorporate latent variables). This does not mean however that SEM is the only possible tool to represent explanations. Simulations have been greatly advanced, as have other representational devices such as FAD (function analysis diagram). The advantage (for the purpose of this paper) of SEM is the ease with which it can extract counterfactuals (by simply assigning a value to a variable and solving the equation). In this way, SEM presents a compact, rigorous, and computationally efficient way to relate to the information of explanations and a direct way to test the set of w-questions answerable by said information. However, as with any modeling practice, it is important to note that there is a certain degree of subjectivity involved when making decisions regarding the abstraction, inclusion (or exclusion), and interpretation of variables.

The value of the proposed procedure is threefold: (i) it gives failure analysts a tool or "procedural check" to critically reflect on some of their own explanatory practices, and (ii) it improves understanding of failure analysis by framing it as an explanatory practice. (iii) It also extends work on the domain of computational argumentation.

The paper is structured as follows. Section "Tools for the procedure outlines the conceptual tools needed for the development and usage of the procedure. Section "The procedure" describes the procedure in detail, and sections "Mechanical Engineering and the virtue of robustness and Civil Engineering and accuracy" provide applications of the procedure in the problem domains of failure analysis in mechanical and civil engineering, respectively. Section "Failure analysis as an explanatory practice" focuses on the value of viewing failure analysis as an explanatory practice and considers its relevance to computational argumentation.

Tools for the procedure

Woodward's counterfactual theory of explanatory power

This paper subscribes to James Woodward's account of explanation where:

> explanation is a matter of exhibiting systematic patterns of counterfactual dependence (…) They do this by enabling us to see how, if these initial conditions had been different or had changed in various ways, various of these alternative possibilities would have been realized instead. Put slightly differently, [generalizations] are such that they can be used to answer a range of counterfactual questions about the conditions under which their explananda would have been different (what-if-things-had-been-different or w-questions, for short). (2003, p. 191)

The structure of an explanation is an argument that employs a generalization in order to track physical dependencies. The quality of an explanation has to do with whether or not it can exhibit *counterfactual* dependency relations. Adequate explanations (according to Woodward) enable this by providing information that can be exploited to answer what-if-things-had-been-different questions (w-questions). Answers to w-questions enable us to see what the outcome would have been if initial conditions had been different. These counterfactual initial conditions oftentimes describe hypothetical situations that would result from changing (intervening on) the values of variables of the explanans. For Woodward, this ability to exhibit systematic patterns of counterfactual dependence is the criterion to establish whether or not an explanation is adequate. Woodward furthermore suggests that the more what-if questions an (adequate) explanation answers, the better it is (Woodward and Hitchcock, 2003).

This paper endorses Woodward's counterfactual theory of explanatory power, but with the following qualification: better explanations are not simply ones that answer more what-if questions, but rather, following Ylikoski and Kuorikoski (2010), better explanations answer more *relevant*[4] what-if questions than alternative explanations.

This criterion for explanatory relevance can be expressed in the following principle: *The quality of an explanation is determined by how many relevant w-questions it can answer.*

Structural equation modeling

SEM was developed in the beginning of the last century by geneticists and economists to combine qualitative cause-effect information with statistical data between variables of interest. It is heavily used in many social sciences because it can model latent variables and errors, while handling many types of relations between variables. A good introduction[5] on how to use and build these equations can be found in Pearl (2009).

Typically, within the SEM literature, a functional causal model (representing an effect as a function of causes and noise) is a set of equations,[6] where some independent variables determine the value of a dependent variable, considering possible errors and/or non-represented factors. It is convention to use the "=" symbol, even though it is an asymmetrical relation that resembles the assignment function ":=". Preferably, each equation in the system should represent what Woodward (2003, p. 328) and Pearl (2009, p. 27) call a (independent) mechanism.

The equations are to be understood as encoding information about how a variable changes if others were to change. By assigning values to the variables, we can understand how a system would (have) behave(d). In the simplest cases, variables are set to 1 if present and 0 if not, and through logical operators they lead to a certain value for the state of the phenomenon under consideration. Additionally, certain quantities (given by the solution of algebraic equations) can activate (or not) some qualitative variables (e.g., turn them into a 1 or a 0).

For instance, consider a toy example of a simple square table with separate legs. The table needs at least three legs to stand, and if it has a weight on it, it can stand as long as the weight is less than a certain amount. Let us idealize and suppose that the weight is always centered, and each leg can hold 5 kg before it breaks. We can model the possible states of the table as follows

[4] Relevancy is the consideration of how important or pertinent the counterfactual is for the aim of the explanation, and while in this account it is considered as a binary, it could be possible to formulate an account where the relevancy of counterfactuals has different degrees.

[5] See also: Hershberger (2003), Hall (2007), Halpern (2008), Halpern and Hitchcock (2011), and Bollen and Pearl (2013).

[6] These can be represented in different ways, for example: $[x_i = f_i(pa_i, u_i), \ i = 1, \ldots n]$ (Pearl, 2009, p. 27), where *pa* are the set of parent variables, and *u* is the error or disturbance due to omitted factors; $[(Y_1 \leftarrow y_1, \ldots., \ Y_k \leftarrow y_k)\varphi]$ (Halpern and Hitchcock, 2011, p. 1113), where Y_i are variables, y_i is the relation with possible variables, and φ is a combination of primitive events; $[Y = aX + U]$ Woodward (2003, p. 327), where X is a direct cause of Y and U is an error term.

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

(where W is the weight (kg) placed on top):

$$\text{Leg}_i = \{0, 1\}\ [\text{for } i = 1, 2, 3, 4]\ (\text{if leg present, } 1,\ \text{if not } 0).$$

$$\text{if}\left(\frac{W}{\text{leg}_1 + \text{leg}_2 + \text{leg}_3 + \text{leg}_4} > 5\right)\ \text{then Leg}_i = 0\ [\text{for } i = 1, 2, 3, 4].$$

$$\begin{aligned}\text{Table\_Stands} = {} & (\text{leg}_1 \vee \text{leg}_2) \wedge (\text{leg}_3 \vee \text{leg}_4) \wedge (\text{leg}_1 \vee \text{leg}_3) \\ & \wedge (\text{leg}_2 \vee \text{leg}_4) \wedge (\text{leg}_1 \vee \text{leg}_4) \wedge (\text{leg}_2 \vee \text{leg}_3) \\ & (\text{if } 1,\ \text{table stands})\end{aligned}$$

Using this scheme, we can explain the current state of the table (e.g., why it is standing or why it collapsed). This is because these equations represent physical dependencies; they represent causal interactions between weights and the breaking of legs. As a consequence, we can answer w-questions by filling-in the variables. For example, if there is no weight and there are three legs (say, the first, the second, and the fourth), then [Table_Stands = 1]. If there is a weight of 6 kg, and 3 legs, then [(6/3 = 2)<5], so [Table_Stands = 1], the table stands. Note that if the weight had been greater than 15, it would not stand (and yet it would, had it had 4 legs with a weight lesser than 20).

The procedure

Building upon the previous ideas, this section offers a 5-step procedure for assessing the adequacy of a failure explanation, for comparing such explanations, and for suggesting possible redesign recommendations. Steps 1 and 2 are instrumental for steps 3, 4, and 5. Sections "Mechanical Engineering and the virtue of robustness and Civil Engineering and accuracy" cover two applications of the procedure. This section focusses on the procedure itself.

1. *Transcription* of the explanation into structural equations (both descriptive text and mathematical equations can be incorporated). This transcription provides a concise summary of the key ingredients of the explanation.
2. *Identification* of the set of desirable w-questions based on the aim of the explanation.[7] For an aim to be fulfilled optimally, the explanation should be able to answer this set of w-questions.
3. *Exploration and corroboration* of which w-questions can actually be formulated and answered in terms of the information offered by the explanation. SEM facilitates this exploration of answerable w-questions by enabling the identification of key epistemic characteristics of the explanation. I call these epistemic characteristics virtues since their presence enables giving answers to sets of w-questions. By identifying these virtues, we can figure out how well an explanation can be exploited to answer the set of w-questions identified in (2), which can be used to assess the performance (and hence quality) of the explanation.
4. *Comparison* between the explanation and a hypothetical counterpart in which a factor or multiple factors of the original explanans are modified. By figuring out whether it is possible to maximize virtues further by tweaking the current structural model, we can either improve the current explanation or show that it is optimized relative to a specific virtue. Given that it is possible to (among other things) add more variables, increase their range, or simplify equations; how do such changes affect the quality of the explanation? This fourth step provides a feedback loop whereby one can improve explanations if they do not provide (all the) relevant answers. Thereby, it allows engineers to engage critically with their own (and others') work. It can also be used to compare competing explanations by contrasting the set of relevant w-questions each can answer.
5. *Validation and recommendations:* Based on previous steps, it is possible to validate an explanation and, in some instances, to provide redesign recommendations.

Glossary

- *Accuracy*: The degree to which relevant (difference-making) factors are captured in a model.
- *Adequacy*: The degree to which an explanation contains the information that can answer the explanatory requirements and aims of a particular audience.
- *Counterfactual*: Hypothetical situation that would obtain if certain conditions were met.
- *Counterfactual dependence*: An event E counterfactually depends on event C if and only if (i) if C had occurred, E would have occurred; (ii) if C had not occurred, E would not have occurred.
- *Explanandum*: What is being explained within an explanation.
- *Explanans*: The model or propositions from which one can derive the explanandum.
- *Robustness*: The range of values that can be given to the variables of a model (while remaining valid).
- *w-questions*: Questions that ask how something would have been different had a different set of conditions been the case.

In the following two sections, this procedure is applied to two case studies on failure analysis, one from Mechanical Engineering and one from Civil Engineering.

Mechanical engineering and the virtue of robustness

Explanation of the failure of a steering shaft and its transcription into structural equations

This section details the case study[8] of a heavy road vehicle's steering shaft rupture, which looks at the investigation aimed at determining whether a fractured steering shaft was the cause or the consequence of an accident (Cleland and Jones, 1996). We can reconstruct the investigation in two steps: first, figuring out (through metallography analysis) whether the shaft had malfunctioned or, alternatively, that the rupture was a consequence of the accident. It turned out that the rupture was an effect of the accident. Given this result, the second step entailed explaining how the accident would transfer an amount of force large enough to

[7] This aim will generally be the aim for which the explanation was created, but it could also be the case that one needs to extrapolate the information from a different explanation to fit the aim of a new explanatory request. In such cases, the procedure operates in the same fashion, but it will likely be the case that optimizing (in step 4) will not be possible without acquiring additional information.

[8] For useful images that illustrate the shaft and the extended explanation, see Cleland and Jones (1996).

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

break an otherwise "healthy" shaft. Their explanation is therefore not concerned with redesigning the shaft, but rather, with showing that the accident caused it to break. This means they take the characteristics of the shaft as fixed, that is, not in need of redesign, which would not happen in a redesign explanation (cf. the explanation of the section "Civil Engineering and accuracy").

The shaft displayed unmistakable signs of shear failure (not only a visual inspection but a scanning electron microscope also confirmed shear failure and fibrous tensile failure at a micro level). Given that the metallography indicated that material properties did not contribute to the failure, they then pursued the second option: the accident had broken the shaft.

Since the question is how the shaft broke, given that the vehicle underwent an accident, and there were unmistakable signs of shear failure, the shaft's failure torque becomes the key variable. They considered the relations between several variables[9] and failure torque. This allows to calculate the maximal force above which a rupture would happen: "the present analysis provides an upper-bound estimate for the failure" (1996, p. 17). Their calculations go as follows:

The equation for the torque required to cause shear fracture of a narrow concentric band is:

$$\mathrm{d}\Gamma = 2\,\pi k_u r^2 \mathrm{d}r,$$

where Γ is the torsional moment, k_u is the ultimate shear stress, and r is the radius.

Ultimate shear stress can be calculated from the empirical expression

$$k_u = \sigma_{\mathrm{TS}}/1.6,$$

where σ_{TS} is the tensile strength, which can be estimated from hardness data through

$$\sigma_{\mathrm{TS}}(\mathrm{MPa}) = 3.2\,\mathrm{HV}.$$

And ultimate shear stress can be formulated as a function of distance to the center,[10] by an empirical equation of the form:

$$k_u(\mathrm{MPa}) = 700 + 0.00878(r/\mathrm{mm})^3.$$

Ultimate shear stress is integrated along the radius (from 0 to 23.5 mm) to get the maximal torque of the core:

$$\Gamma = 2\pi \int_0^{23.5} (700 + 0.00878 \cdot r^3) \cdot r^2 \cdot \mathrm{d}r = 20.6\,\mathrm{MN\,mm}.$$

The shaft's case can be calculated by the formula of a cross-section:

$$\Gamma = \frac{2 \cdot \pi \cdot\ k_u \cdot (r_1^3 - r_2^3)}{3}.$$

Which gives a torque of 9.8 MN mm for the case. The total maximal torque is then 30.4 MN mm (9.8 + 20.6).

From here, it follows that

$$F_{\max}{*}\text{length of steering arm (250 mm)} \\ = \text{total torque(30.4 MN mm)}.$$

Giving a maximum force of approximately 0.1216 MN, which is equivalent to 12.4 tf. This means that a force greater than 12.4 tf would likely break the shaft. The paper follows with the observation that a vehicle of 20 t with a (conservative) deceleration of 5 g has a collision force 100 tf (8 times more than is needed to break the shaft). In other words, the shaft broke as a result of the collision.

The authors conclude "The steering shaft had been subjected to a large axial torque, sufficient to cause gross yielding of the cross-section and fracture by a ductile mechanism. There were no indications that failure was promoted by prior defects or inadequate mechanical properties. If a small fraction of the likely collision force had been applied to the end of the steering arm, this would have been sufficient to cause failure. We therefore conclude that the failure was a consequence of the accident, and not its cause" (p. 18).

The explanation given by the authors can be regimented further by executing the first step in the procedure, namely the *transcription* of the variables and their relations into structural equations (which the engineers themselves did not provide) which can be used to set counterfactual scenarios by filing in the variables:

$$\text{Total torque} = \text{core torque}\left[\int_0^{r_2} (2\pi \cdot\ k_u \cdot r^2 \cdot\ \mathrm{d}r)\right] \\ + \text{case torque } [2\pi \cdot k_u(r_1^3 - r_2^3)/3]$$

$$\text{If (Force} > \frac{\text{Total torque}}{\text{arm length}} + U\text{) then Shaft\_Break} = 1\text{; else} = 0$$

where U is error and omitted factors.

These structural equations provide a concise representation of the key ingredients of the explanation, highlighting what is epistemically relevant. It starts from the explanation (i.e., it assumes the validity of the explanation) in order to condense the relations that are of interest.

Identification of relevant w-questions

Moving to the second step of the procedure, we now need to identify the set of w-questions that (when answered) would satisfy the aim of the explanation. The original aim was to figure out whether the broken shaft caused or was a caused by the accident. An explanation at the very least should clarify this; however, once this minimum requirement is met, we can inquire whether it fulfills its aim optimally. To do so, it would need to satisfy as many

[9] For the shaft's core: yield stress (736 MPa minimum), which is the limit of elastic behavior. Tensile strength (1079–1324 MPa), which is the capacity of a material to resist elongation. Elongation (8% minimum), impact energy (59 Jcm-2 minimum). The shaft's case had a required Vickers hardness of 600–780 HV (but was measured to be 880), which is a materials' ability to resist plastic deformation determined by the load over the surface area of an indentation. Predicted hardness of the core was 340 HV and was measured to be 350 HV at the center, growing up to 400 (at a radius 22.5 mm). The radius of the shaft was 23.5 mm core and 1.5 mm of (concentric band) casing. The length of steering arm was 250 mm. No details are given about other possible variables that could relate (such as contracted distance of the vehicle).

[10] This results from having calculated HV at different concentric points.

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

relevant w-questions as possible. This means it should give us as much information as possible about when the collision would have caused the shaft to break and when a broken shaft would have caused an accident. The more scenarios (answers to w-questions) in which we know whether the shaft breaks or not, the better our explanation becomes. An example of such a relevant w-question would be "what-would-have-happened at half the collision force?". More concretely, only the counterfactuals that consider changes in force are relevant.

If the aim had been to redesign the shaft, other counterfactuals would become relevant, such as ones considering how changes in the shaft's characteristics would affect the outcome, for example "what-would-have-happened if the total radius were increased by 2 mm". Even though these questions provide redesign information, they are not relevant for the explanatory aim in the case at hand, which was to figure out whether the collision broke the actual shaft.

Exploration and corroboration of which w-questions can be answered

We need an explanation that can answer the set of w-questions related to different force values ("what-would-have-happened-if force was *n*"). The third step in the procedure is to figure out whether our explanation can indeed provide answers to this set of w-questions. The structural equations presented in the section "Explanation of the failure of a steering shaft and its transcription into structural equations" can help us answer many of the w-questions that can be formulated. By plugging values into the structural equations, we can obtain answers. For example, "what-would-have-happened-if the impact force = 2 tf", where the answer would be 2 tf <12 tf, so Shaft_break = 0. We could also answer several *irrelevant* w-questions, that could nonetheless be interesting for other purposes (such as redesign). For instance, "what-would-have-happened if the arm length was 260 mm?". If the arm length was 260 mm, then 100 tf > (30.4 MN/0.26 m), so Shaft_break = 1. This tells us that modifying the shaft length to 260 mm would not have prevented it from breaking. However, given the aim, we are only interested in w-questions about different force values.[11]

If the variable of force in our structural equations can capture a greater range of values, we will have more information about the conditions in which the shaft breaks or not (we will be able to specify a greater number of instances). Looking at the range of values force can take (while the model still outputs valid results), we can assess the number of scenarios that can be captured and hence how many w-questions can be answered. If a variable can take a set of integers ranging from 1 to 20, it will enable capturing more scenarios than a variable that can only be satisfied with a set of integers ranging from 2 to 5. This amount of range will be called *robustness*,[12] which refers to how much change we can introduce in the variables of the equations of a model (or its assumptions) while keeping its results valid. This means that the greater the range the variables have, the more robust our model is; and by extension, the more robust the model, the greater the range of values its variables can capture (in this particular case, force).

By looking at the robustness of the model, we can learn about the size of the set of answerable w-questions, because the greater the range of force, the more w-questions (relevant to the aim) are answerable. Looking at the structural equations, we can see that the model is not well defined when Force $\approx$ (Total torque/arm length). When these two values are close, the result is uncertain (just how much will depend on U – the error). The model is valid in ranges where Force $\neq$ (Total torque/arm length), so it is rather robust.[13] The model can tell us what happens to the explanandum when the variables take values in the ranges of [0, 12.3 tf) and (12.5 tf, ∞). If this range were shrunken, we would be able to capture fewer scenarios (see section "Validation and recommendations", Figure 1 for a visual representation of the ranges), in the sense that the subset of scenarios would decrease (if $U_1 > U_2$, then the set of captured scenarios by the equations with error-term U_1 is a proper subset of the captured scenarios with error-term U_2. In other words, the set of scenarios captured – in this case – is inversely proportional to the error). By extension, if we could establish with precision what happens near the 12.4 range, we would increase the number of scenarios. In other words, an increase of robustness would increase the number of answerable (and relevant) w-questions. In explanations where the aim is similar to the current example, robustness is a good indicator of how good the explanation is. From this, we can assert that the explanation given by the engineers was good, since it does indeed answer a big array of relevant w-questions, which derives from the fact that their model is robust (it is possible to draw conclusions from substituting F with many values).

Counterfactual comparison

The third step shows that our explanation is good (it satisfies its aim). The question now becomes whether it can be improved and how well it compares with others, which leads us to the

[11] By focussing on the set of relevant w-questions, we dramatically narrow down the space of w-questions, but one might wonder whether there could still be a combinatorial explosion of relevant w-questions in complex cases (e.g., considering parameters that might have to account for unforeseen situations).

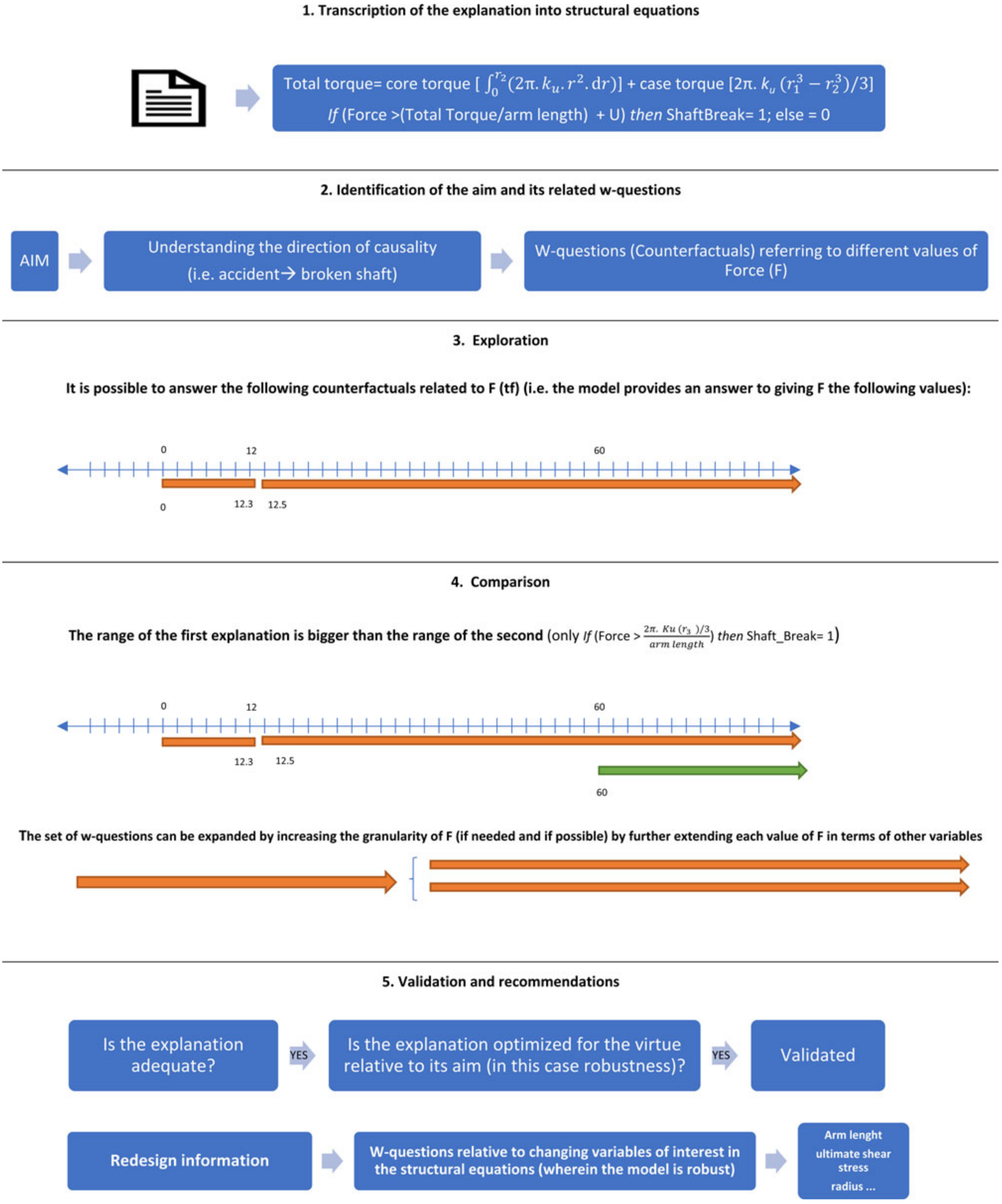
It is important to note that checking whether a model can answer a set of w-questions and answering said w-questions are different operations. While it is true that exhaustively checking the latter might lead to a combinatorial explosion, the former is normally contained. The main reason is that usually each structural equation represents a generalization, and each equation can be linked to a set of answerable w-questions in a linear manner. Put differently, sets of answerable w-questions are normally limited by the number of structural equations. Therefore, the validation and optimization of the model will typically not lead to a combinatorial explosion.

There are however cases where it might be necessary to consider several solutions of w-questions in combination, such as when using a model to give precise redesign recommendations. One example could be trying to improve a design by simultaneously considering several modifications to different parameters in order to figure out their optimal values. In such instances, the complexity might grow exponentially. For such cases, it is possible to apply meta-heuristic optimization strategies, such as the use of genetic algorithms, where the fitness function would consist of a selection of the model's structural equations. In the example of the shaft, we could consider the first equation, where we would like to maximize maximum sheer stress (total torque) before rupture. We could then specify certain constraints (upper-lower values and the size of step increments) of possible radii, tensile strengths of known materials, etc., and try to find optimized values for these parameters.

The advantage of the procedure in such applications is that the fitness function will be easy to specify given the translation into structural equations. However, its usefulness will depend on the ability to constraint the search space to reflect external considerations (e.g., in our shaft example, there are limitations on the length of the shaft based on the overall vehicle design which are not captured by the model).

[12] Typically, robustness refers to insensitivity to variation. To capture this idea, the notion of range is used (the more range variables can take, the less sensitive the model is to variations).

[13] Negative values of F would have no meaningful interpretation as in this particular case we are only interested in the modulus of the force and whether it exceeds a threshold value (the direction of rotation of the shaft does not make a difference).



1. Transcription of the explanation into structural equations

Total torque= core torque [$\int_0^{r_2}(2\pi . k_u . r^2 . dr)$] + case torque [$2\pi . k_u\ (r_1^3 - r_2^3)/3$]

If (Force >(Total Torque/arm length) + U) *then* ShaftBreak= 1; else = 0

2. Identification of the aim and its related w-questions

AIM

Understanding the direction of causality (i.e. accident→ broken shaft)

W-questions (Counterfactuals) referring to different values of Force (F)

3. Exploration

It is possible to answer the following counterfactuals related to F (tf) (i.e. the model provides an answer to giving F the following values):

4. Comparison

The range of the first explanation is bigger than the range of the second (only *If* (Force > $\frac{2\pi .\ Ku\ (r_3\ )/3}{arm\ length}$) *then* Shaft_Break= 1)

The set of w-questions can be expanded by increasing the granularity of F (if needed and if possible) by further extending each value of F in terms of other variables

5. Validation and recommendations

Is the explanation adequate?

Is the explanation optimized for the virtue relative to its aim (in this case robustness)?

Validated

Redesign information

W-questions relative to changing variables of interest in the structural equations (wherein the model is robust)

Arm lenght
ultimate shear stress
radius ...

Fig. 1. Summary of the procedure applied to the case of a broken shaft.

fourth step: comparing our current explanation with possible hypothetical counterparts (where some factors are modified) to figure out whether the explanation can be optimized.

A model that maximizes the value range of force (enabling answering more relevant w-questions) is better; therefore, it is an advantage to have a maximally robust model. The question arises whether we can tweak the model to improve robustness further. There are several ways to do this. Theoretically, we could increase robustness by specifying more carefully the behavior of the system when the force approaches the breaking point

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

(12.4 tf). But it is likely that improvements along those lines will come from improving external factors to the model, such as measuring capabilities. Another option is to try to increase the information relative to F. If each value of F captures a tuple of other values, its range (arguably) increases. Put differently, we can also increase robustness by enhancing the granularity of F, which can be done by expressing F in terms of new variables. If we expand F into the expression Kinetic energy (K)/distance (d), we cover more scenarios,[14] as each value of F can be obtained by a set of $[K/d]$ pairs. By specifying $F = (K/d) = ((m \cdot v^2)/(2 \cdot d))$ we can gain information about how the speed might affect the shaft failure (given the mass "m" of the vehicle and a deformation distance "d"). The distance will depend on the rigidity of the vehicle, and the velocity will depend on how fast it was going before impact. The reason expanding F provides more information is that we would know more about how different speeds might affect the breaking of the shaft. However, this is only useful if we can actually come to know the values of "d" and "m". Including unknowable variables cannot improve the quality of the explanation, since one cannot set counterfactuals without knowing the value of said variables. Knowing which variables can be determined informs step 4 by telling us what modifications are sensible and which are not.

Increasing the granularity of F will improve the explanation, but only if this increase can be actualized (if we can determine and measure the variables that account for F). Such a decision needs to be considered by the engineering team working on the specific case. Ultimately the team must decide whether there are other (measurable) variables that can enhance granularity. However, should there be such a possibility, it is wise to implement it since it would improve the explanation.

Furthermore, we can illustrate how modifications that diminish robustness are detrimental to the quality of the explanation:

> Suppose a hypothetical explanation of the same failure as follows. We know that the case's tensile strength is much higher than the interior, so we can set an unreasonably conservative upper bound limit by just taking $k_u = 1760$ MPa for all the shaft (instead of only the case), and calculate a torque that would certainly break the shaft by the following equation:
>
> $$\Gamma \approx \frac{2 \cdot \pi \cdot k_u \cdot r^3}{3} \approx 16.3\ \text{MN mm}.$$
>
> This means that given an arm length of 250 mm, a force greater than 60 tf would surely break the shaft. Using the same back of an envelope calculation, the engineers used (20 t * 5 g of deceleration = 100 tf of impact force), we can explain how the shaft broke: 100 is much greater than 60 (which is already greater than the real limit). This represented by the following structural equation:
>
> $$\text{If } \left(\text{Force} > \frac{2\pi \cdot k_u\ (r^3)/3}{\text{arm length}}\right),\ \text{then Shaft\_Break} = 1.$$
>
> While this also explains the failure, it seems of an inferior quality.[15] The first thing to notice is that robustness is reduced when using only the equation of the case instead of a more precise model that uses the equation of the case and the equation of the core (see Fig. 1 for a visual representation). All the scenarios that our hypothetical explanation captures can also be captured by the engineer's explanation, but their explanation also captures ones where the collision force is smaller. Put differently, our hypothetical explanation tells us that the shaft breaks if force is greater than 60 tf but cannot tell us much about numbers below 60. The engineer's explanation tells us what happens if the force is greater than 12.4 tf. This means that it captures a greater range (all the values between 12.4 and 60). Additionally, the engineer's explanation can also account for what happens below 12.4.

Validation and recommendations

Based on the results of steps 3 and 4, we can validate the explanation (robustness is maximized). If we applied the procedure to the hypothetical explanation from the previous section, it would not be validated, since its robustness could be improved. The illustration below summarizes the process (Fig. 1). The present analysis provides a means to engage with the explanatory content by considering how well suited it is to answer relevant w-questions and whether it can be improved. It, therefore, provides tools to gain understanding as to why the available explanation is adequate or not.

Consider, by mode of contrast, that the aim of the explanation had been to adjudicate responsibility for the accident. In such a case, the relevant w-questions would not be answered by the current explanation (and robustness would not be the virtue needing optimization). In such a scenario, an adequate explanation should have included other factors (e.g., the model would include variables of actors which could be set to 0 or 1 to check whether their involvement was conducive to the accident or not). This information is not present in the explanation (a fact that is easily observable by looking at the structural equations) and hence it is not adequate for the aim of adjudicating responsibility.

Given that the main aim was not redesign, the procedure need not be used to give any recommendations. Nevertheless, it would be possible to use the structural equations to question what the result would had been with a modified design. For example, we could change the values of length of the shaft and check if it would have broken under 12.4 tf. Other relations between variables can also be established by using the equations (i.e., by inputting certain values into variables of interest one can obtain the value of other variables). We now turn to a case in which redesign was the main aim of the explanation.

Civil engineering and accuracy

Collapse in a convention center and its transcription into structural equations

This section details an explanation given by the independent firm Wiss, Janney, Elstner Associates (WJE), hired by the owners of the David L. Lawrence Convention Centre to explain the collapse of an expansion joint in their convention center. The final aim of WJE's explanation was to redesign (and fix) the failed expansion, which was implemented by Thornton Tomasetti (a hire of the original architect) under their supervision.[16]

[14] Note that there are other ways to calculate F. For example, $F = v^*m/t$, where v is initial velocity and t is time the collision lasts.

[15] We could also provide an extremely lower bound limit (by considering the whole shaft as having the average tensile strength of the core), but the resulting explanation would still be of an inferior quality.

[16] The brochure of the services they provided reads: "After determining the cause of the loading dock failure, WJE assessed the conditions of structural systems throughout the rest of the building and recommended retrofitting all beam end connections like the one that failed in order to prevent future collapses. To further assist the SEA, WJE reviewed the engineer of record's designs for reconstruction of the collapsed area and provided quality assurance services during the reconstruction and the installation of the retrofit. Finally, WJE provided litigation consulting to SEA" (https://www.wje.com/projects/detail/david-l-lawrence-convention-center).

The 4-story convention center was rebuilt in 2003. An expansion joint (with 25 slots) split the center in two. Joints are often one of the weakest points in structures. Their function is to connect while absorbing tension, temperature-induced size changes of the connected parts, vibration, etc. (Delatte, 2009). In 2007, a tractor-trailer parked on the second floor collapsed a concrete slab. The failure had occurred at an expansion joint exposed to an ambient temperature of −19°C. The colder a joint is, the more open it is, making it more likely to collapse under weight. Upon investigating the incident, it came to light that in 2005 a similar failure had occurred, resulting in a beam dropping onto a column. At the time, this incident had not been disclosed to the relevant authorities and was not considered relevant.

The reconstruction by WJE of the details goes as follows:

Temperature-induced displacement can be calculated by[17]: $\delta T = \alpha (\Delta T) L$.
Additionally, the displacement caused by load can be calculated by[18]: $\delta = PL/AE$.
And if thermal deformation is restrained, the force build up is: $P = \alpha(\Delta T)AE$.

They estimated the amount of free movement required to be 41 mm, assuming a temperature change of 28°C. This is because the α of steel is 10^{-5} mm/mm/°C and the (half) length of the building was 133 m. A finite element analysis showed that the distortion (of the high-strength steel angles) generated 630 kN of tension at the connection welds, with 8 mm of displacement. It was also noted that "[w]ith lower strength A36 steel, the force on the welds would have been reduced by 40%" (Delatte, 2009, p. 210).

Slots should be long and loose enough, and the bolts centered, allowing free movement that would not lock the joint or bear against the edge. Other possible factors for locking the joint might be corrosion, paint, and debris.[19] These, among other problems, are why the *Manual of Steel Construction* of the American Institution for Steel Construction suggests either double line of structural columns or low friction sliding connections for this type of joint (1998, part 8).

To make matters worse, the slots were welded only in the outer edges, making them weaker. When the tension grew too strong, they simply pulled free. The main investigation (WJE, 2008) concluded:

> "The main design issue was that the slotted hole expansion joint was almost guaranteed to fail because of significant friction and insufficient room for thermal contraction […] Other design errors did not contribute significantly to the collapse; these errors included inadequate length of the slot and no limitation on bolt torque. Materials and fabrication issues included steel with too high a strength-ASTM 92, not 36. This high strength kept the angles from bending and caused them instead to tear away at the weld […] there was little evidence of the bolts actually sliding within the slot; instead, the threads seem to have worn away at the same spot." (2009:209)

The fix consisted in welding 1-foot-square steel seats with Teflon pads underneath the joints of the replacement beam (and 25 others). This structural design element was indicated in early drawings but never executed (Rosenblum, 2007).

The explanation can be further regimented by executing the procedure's first step:

If (Displacement > Room_Available), *then* Not_Enough_Room = 1.
If (Total_Tension > Welding_Resistance − U), *then* Too_Much_Tension = 1.
Pulls_Apart = Not_Enough_Room ^ Too_Much_Tension.

*Where Displacement is calculated by: $\delta T = \alpha\ (\Delta T)\ L$.

*Room_Available is a fixed design feature.

*Total_Tension is calculated by the tension generated from thermal change: $P = \alpha\ (\Delta T)\ A\ E$ plus any weight there might be on top (such as a tractor trailer): $P = m \cdot g \cos\theta$.

*And Resistance is a feature of the material where U accounts for imprecise factors such as corrosion, paint, debris, and maintenance.

Note that for Pulls_Apart to be 1 you need both conditions (if there is enough room, there will not be too much tension, if there is no room but not enough tension, it will not pull apart). These structural equations provide a concise representation of the key components (ingredients) of the explanation, highlighting what is epistemically relevant: the relation between variables (such as displacement and room available) and their connection to whether the joint fails or not.

Identification of relevant w-questions

In the second step of the procedure, we need to identify the set of w-questions that (when answered) would satisfy the aim. The aim of the explanation was to offer information that could be profitably used to redesign the malfunctioning element(s). This means that we need to answer the set of questions that tell us whether certain modifications of the current design would prevent future failure. We need information about the conditions in which the inclusion, exclusion, or modification of certain factors of the joint prevents failure. The aim of redesign is maximally fulfilled when the explanation can answer numerous w-questions that relate to how changes might prevent the future failure of the joint. An example of such a question would be "what-would-have-happened if instead of only steel ASTM A92 slots the joint would have used steel ASTM A36 and a Teflon supporting bracket underneath?" Note however, as with the previous case, that not all w-question are relevant. For example, "what-would-have-happened if the temperature the night of the accident was 0°C?" is not a relevant w-question. While it might tell us why it collapsed the day it did and not earlier, it does not give us any information about how certain modifications of the joint would prevent it from failing. This contrasts with the previous case study, where we kept the shaft structure fixed and considered how changes in force would or would not break it. Here, we consider the temperature drop fixed (to a certain extent) and consider how changes in the design would have averted the collapse. This is not to say that relevant questions should not consider worst-case scenarios in the background (such as even more dramatic temperature drops), but the focus should not be on tracking the dependencies between factors external to the design (like temperature) and joint failure. The focus should be on modifications of the design and how they

[17] Where δT is the amount of displacement, α is the thermal expansion coefficient, ΔT is the change in temperature, and L is the length.

[18] Where P = force, A = cross-sectional area, and E = the modulus of elasticity.

[19] This means that in practice this type of system is not a good idea. In fact, an acclaimed engineer criticized this way of doing joints in the first place: "I've only seen the slotted hole connection used one other place in an expansion joint in 30 years of doing engineering. And it fell in that place, as well" (Houser and Ritchie, 2007).

affect the outcome (while keeping in mind possible roles played by external factors).

If the aim were (e.g.) to figure out why it collapsed when it did and not a day before, other counterfactuals would be relevant, such as ones considering variations in temperature and how they affect the success or failure of the joint (captured by 0,1 values of "Pulls_Apart"). But the aim is to redesign. In terms of our structural equations, the modifications that are informative are ones that modify the value of the variables "Welding_Resistance" and "Room_Available" (which could be further cashed out in terms of materials and dimensions if need be).

Exploration and corroboration of which w-questions can be answered

We need an explanation that can answer the set of w-questions related to modifying the current design. The third step in the procedure is to determine whether the explanation can indeed provide answers to this set of w-questions. The structural equations presented in the section "Collapse in a convention center and its transcription into structural equations" can help us answer many of the w-questions that can be formulated. By plugging values into the structural equations, we can obtain answers. For example, "what would have happened if instead of ASTM 92 steel the brackets would have used steel ASTM A36 and incorporated a Teflon coating low friction bracket?" The tension would be reduced dramatically, meaning Total_Tension < Welding_Resistance, So Pulls_Apart = 0. (It would also be possible to give these details in a fine-grained manner calculating relative values of tension and so on, but we choose latent variables to make the argument easier to read).

We could also answer several irrelevant w-questions. For instance, "what-would-have-happened if the maximum temperature was 28°C and the minimum 25°C?" Then Displacement would be: $\alpha\ (\Delta T)\ L \approx 4$ mm, so Displacement < Room_Available, which means Pulls_Apart = 0. This would explain why it did not collapse on a given day (with those temperatures). However, even though the result is favorable (the joint does not collapse), this is not an interesting question (for our aim): we cannot intervene in order to modify temperature so as to avoid future collapses.

The quality of the explanation hinges on how well it answers *relevant* w-questions. In order to provide adequate information for answering such questions, the explanation must contain a detailed representation of factors that make (or could make) a difference to the outcome (failure or success of the joint). To see why, consider how if one ignores an operative factor, it might end up causing a failure in the future. In fact, this is what happened in 2005, when an incorrect evaluation of the difference making factors resulted in inaccurate engineering: not adding a low-friction support bracket ultimately led to the failure in 2007.

The importance of capturing elements that make a difference can be seen by how the engineers consider different factors as being causally relevant or not. For example, they consider insufficient room for thermal contraction as a difference maker, but they also consider other factors which were not difference makers, such as the length of the slot, limitation on bolt torque, whether the bolt was centered, paint, corrosion, and debris.

The greater the number of (causal) factors (i.e., factors that make a difference) that are identified, the better our explanation becomes, since it leaves fewer important things out. This virtue is called *accuracy*. By looking at the accuracy of the model, we can learn about the size of the set of answerable w-questions, because the greater the number of operative factors present in the explanation, the more w-questions (relevant to the aim) are answerable.

The number of relevant factors can be accounted for by looking at the structural equations, since they capture the factors that are deployed in the explanation. If we have a greater number of variables in our structural equations, we can capture more scenarios, hence we will have more information about the conditions in which modifying the joint in different ways leads to a collapse or not (we will be able to specify a greater number of relevant instances).

An increase in accuracy would increase the number of answerable (relevant) w-questions (having more factors enables formulating a greater number of counterfactuals). Furthermore, it is because the explanation is accurate that a considerable number of w-questions can be answered. This means that in explanations where the aim is similar to this one (redesign), accuracy is a good indicator of the quality of the explanation.

WJS's explanation does indeed answer a big array of relevant w-questions, because their model is accurate (it contains several variables that make a difference and considers many possible candidates that upon further evaluation are not considered to make a difference).

The accuracy of their explanation results in the possibility of repairing the failure (redesigning), which was the main aim of the explanation: "A more reliable detail for this type of connection is a low-friction supporting bracket (…) the bolts were removed and Teflon-coated supporting seats were added" (Delatte, 2009, p. 210). The recommendations (and actions) to improve the problem were based on the accuracy of the explanation, namely the need to lower friction and augment room for displacement. Furthermore, the accuracy of the explanation allows to know how the fixes should be implemented (e.g., the dimensions of the supporting seats will depend, among other things, on the expected thermal expansion).

Counterfactual comparison

The third step shows that the explanation is good. The question now becomes whether it can be improved and how it might compare to other alternatives, which leads us to the fourth step: comparing the current explanation with possible hypothetical counterparts (where some factors are modified).

The failure analysts aimed for accuracy in their investigations as can be seen in their quite thorough consideration of what factors played a role and which ones did not contribute. What needs to be analyzed is whether accuracy can be further improved.

The decision as to whether all relevant bases were covered (i.e., accuracy was optimized) relies on the team of engineers with access to the whole information about the case. However, what can be said is the following: if the aim of the explanation is redesign, should there be a factor that affects the displacement, it needs to be included in the explanation in order to make the explanation more accurate (and the redesign more effective). What the procedure can establish is how many factors are indeed included in the model and compare competing explanations to show that the one that is more accurate will perform better. To see why, we can use a hypothetical explanation that is less accurate.

Contrast the engineers' explanation with an (worse for this aim) explanation that would pursue simplicity. To do so, one could for instance remove parts of the structural equations or

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

leave certain details of the explanation out. Consider ignoring the first equation, that refers to having enough room. In theory, the failure could just be accounted for by simply referring to the tension overload. It would be possible to build an explanation that only focussed on the tension generated in the slots, without paying attention to the available room. For example:

If (Total_Tension > Welding_Resistance − U), *then* Pulls_Apart = 1; else = 0.

This would in theory explain the failure: if the tension is greater than it can hold, it will break. It does tell us to some extent that if the tension had been smaller or the welding resistance larger it would not have broken. However, this is very limited. The simplicity of the explanation could in some cases benefit the explanatory aims of certain audiences (e.g., in a court room or a classroom), but it certainly would not aid the implementation of sound changes to the construction.

Our hypothetical explanation misses the point, since it would only recommend changing the material of the bolts, which would not address the underlying issue of thermal expansion (leading to possible complications down the line). This thought experiment highlights that in order to have an explanation that provides the right sort of information toward redesign, one needs accuracy.

We could also compare it to the explanation of 2005 that did not address the problem in the first place:

> Another beam in the convention center had caused a problem in 2005. The beam dropped 2.5 inches, said Mary Conturo, executive director of the Allegheny County Sports & Exhibition Authority. “At that time we called in anyone that was responsible for that—the architect, the structural engineer, the steel fabricator and the steel erector,” Conturo said. “They all came and reviewed the situation and their response was that there was a bolt that was too tight, and that other bolts were checked, and that was the extent of what was done at that time.” (Engineering News-Record, February 22, 2006)

Had their explanation been more accurate (i.e., identified the factors that would cause a problem later on) and had they acted upon the knowledge it would have provided, it would be fair to say that there would not have been a collapse in 2007. If they had not only considered the tension but also the room required for thermal expansion, they would likely have re-evaluated the design as needing a support bracket (as was done after the collapse in 2007).

Validation and redesign recommendations

The results of step 3 show that the explanation is adequate (it is a good explanation). The results of step 4 show that it is optimal for the aim and is better than the alternatives considered. Given that accuracy is maximized in the engineer’s explanation, it can be validated. We can easily see that both our hypothetical explanation and the explanation given in 2005 are not validated (a visual summary can be found in Fig. 2). The procedure, therefore, delivers insight as to why certain explanations are adequate and others are not, while allowing to improve those which are not. Similarly, it provides resources to justify the preference of certain explanations over others, while encouraging critical reflection on the results of failure analysis.

Furthermore, the procedure can be used to provide information useful toward redesign. Accuracy is key to give the right recommendations (in this case, the addition of a low friction supporting seat); but it is also important for the correct implementation of the recommendations. Once the solution of a low friction supporting seat is proposed, there are other things to consider such as what materials to use (e.g., Teflon bearing vs. other elastomeric pads), the dimensions of the support bracket, etc. All these details need to be adapted to the case at hand, meaning that the implementation of the solution will benefit from an improved accuracy that takes into account the specific details and factors of the explanation. The information encoded through the procedure allows to make inferences about what type of redesign implementations are reasonable. For example, to know the dimensions of the supporting brackets, one must know not only the dimensions of the joint, but also how much it is expected to expand in unfavorable weather conditions. Using the structural equations, we can plug in such unfavorable conditions (e.g., a 40° drop in temperature) to see what the expansion would be, and by extension how large the support should be. A similar argument can be made for the resistance of the material needed for the supporting seat.

Failure analysis as an explanatory practice

This section addresses the aim of improving understanding of failure analysis in light of the procedure. Given the fundamental qualities of the outcome of failure analysis (e.g., they can be compared), I argue that failure analysis is best understood as an explanatory practice (a practice that produces explanations) rather than a cause-finding practice.

Failure analysis is often labeled as a practice of *cause finding*. Several authoritative definitions of failure analysis share this view, for example, take the definitions provided by the ASM international handbook, the McGraw-Hill *Electronic Failure Analysis Handbook*, or by the fourth edition of Elsevier’s *Machinery Failure Analysis and Troubleshooting*:

> [F]ailure analysis is a process performed in order to determine the causes or factors that have led to an undesired loss of functionality. (Becker and Shipley, 2002, p. 315, emphasis added)

> Failure analysis is the process of determining the cause of failure, collecting and analyzing data, and developing conclusions to eliminate the failure mechanism causing specific device or system failures. (Martin, 1999, p. 1, emphasis added)

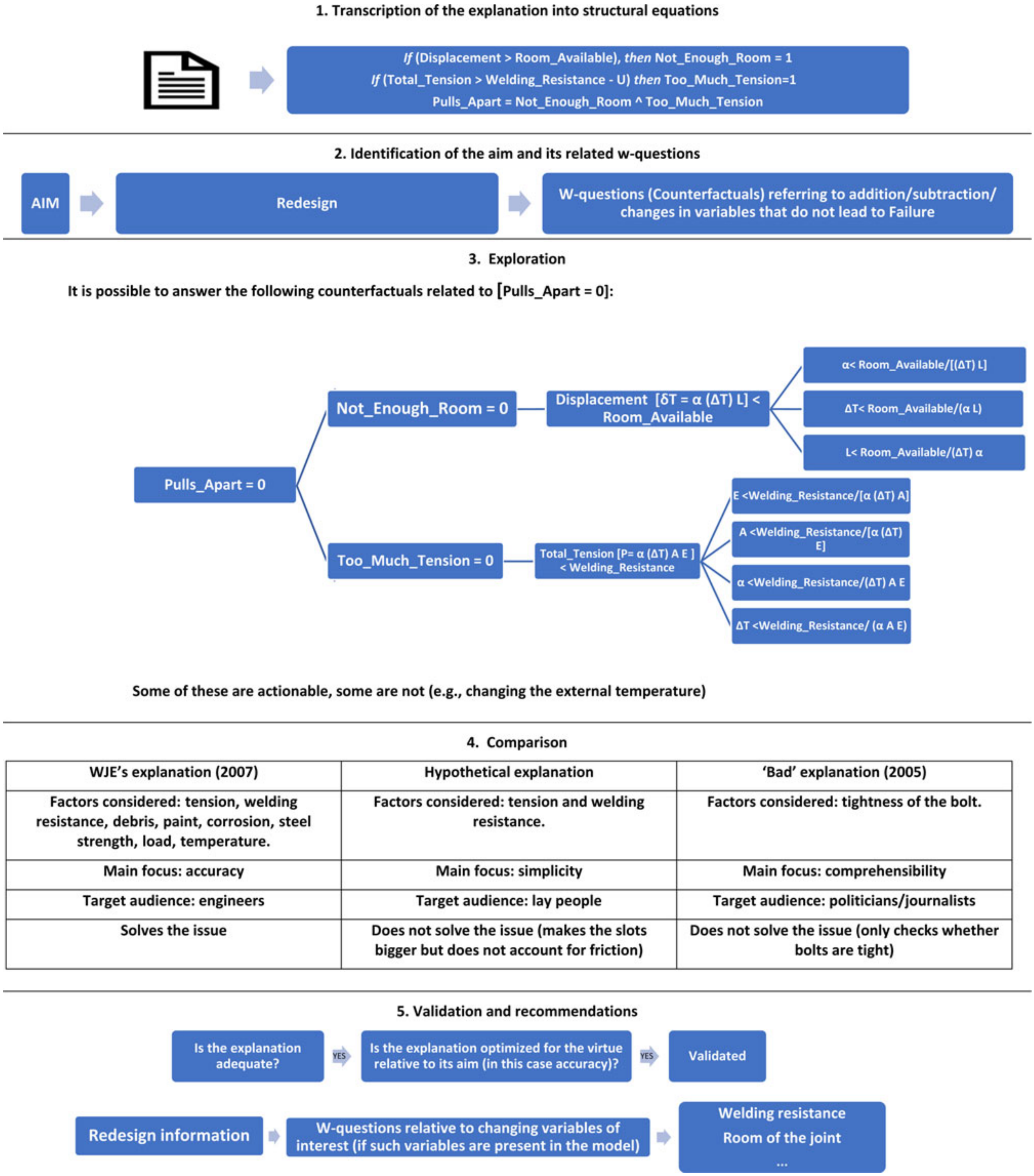
> [The job of the failure analysist] is to define the root cause of the failure incident and to come up with a corrective or preventive action. (Bloch and Geitner, 2012, p. 15, emphasis added)

Many similar definitions are found throughout the literature, sometimes preferring the term “factor” or using qualifying adjectives (e.g., primary, immediate, direct, underlying, probable, latent, secondary, or root cause). These adjectives are sometimes trying to capture the multifaceted nature of failure analysis. After all, failure analysis is a rubric used to designate a variety of different practices and methods (FMEAs, Fault Tree Analysis, Root-cause Analysis, What-if Analysis …)

I contend that framing failure analysis as “cause finding” does not do justice to what the practice is really doing. It neglects to acknowledge that failure analysis is concerned with explaining.

This is especially relevant for redesign. Finding a cause does not give the necessary information for redesign; one needs to know how this cause connects to everything else to understand (to explain) the reasons for the failure in order to avoid it or to learn from it. Failure

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press


1. Transcription of the explanation into structural equations
If (Displacement > Room_Available), *then* Not_Enough_Room = 1
If (Total_Tension > Welding_Resistance - U) *then* Too_Much_Tension=1
Pulls_Apart = Not_Enough_Room ^ Too_Much_Tension
2. Identification of the aim and its related w-questions
AIM
Redesign
W-questions (Counterfactuals) referring to addition/subtraction/ changes in variables that do not lead to Failure
3. Exploration
It is possible to answer the following counterfactuals related to [Pulls_Apart = 0]:
Pulls_Apart = 0
Not_Enough_Room = 0
Displacement [δT = α (ΔT) L] < Room_Available
α< Room_Available/[(ΔT) L]
ΔT< Room_Available/(α L)
L< Room_Available/(ΔT) α
Too_Much_Tension = 0
Total_Tension [P= α (ΔT) A E] < Welding_Resistance
E <Welding_Resistance/[α (ΔT) A]
A <Welding_Resistance/[α (ΔT) E]
α <Welding_Resistance/(ΔT) A E
ΔT <Welding_Resistance/ (α A E)
Some of these are actionable, some are not (e.g., changing the external temperature)
4. Comparison

WJE's explanation (2007)	Hypothetical explanation	'Bad' explanation (2005)
Factors considered: tension, welding resistance, debris, paint, corrosion, steel strength, load, temperature.	Factors considered: tension and welding resistance.	Factors considered: tightness of the bolt.
Main focus: accuracy	Main focus: simplicity	Main focus: comprehensibility
Target audience: engineers	Target audience: lay people	Target audience: politicians/journalists
Solves the issue	Does not solve the issue (makes the slots bigger but does not account for friction)	Does not solve the issue (only checks whether bolts are tight)

5. Validation and recommendations
Is the explanation adequate?
YES
Is the explanation optimized for the virtue relative to its aim (in this case accuracy)?
YES
Validated
Redesign information
W-questions relative to changing variables of interest (if such variables are present in the model)
Welding resistance
Room of the joint
...

Fig. 2. Summary of the procedure applied to the case of the convention center.

analysts indeed provide such explanations, hence they are not simply doing "cause finding". Finding causes is only part of an explanation, but of similar importance is deciding which causes should be included or excluded, building a coherent story about how these causes interact, separating them into causal factors and background conditions (and understanding how these background conditions might be affected in abnormal circumstances). After all, one can trace back the "causes" of an event pretty far and wide, which is why explanations conceal such a causal history by highlighting the relevant factors and making salient that which is pertinent to the goals of the explainer (e.g., redesigning a failed component).

Furthermore, explanations can be compared to each other, and can be improved, which is not always true for causes. The term "cause" often can also be problematic. For example, many analyses try to establish the root cause. What exactly the root cause is will be highly context-dependent, which might lead to confusing results (e.g., exactly when should one conclude the "root" has been reached?). Bhaumik (2009) shows how the proper root

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

cause is hardly ever found, and that what is often called "root cause" is merely the primary cause of failure or a simple physical cause. Similarly, causes can be a variety of things: events, objects, processes, and even absences (such as when we say that the cause of an accident was the omission of safety procedures, or that the failure to step on the brakes on time caused a crash). How all this variety of types of causes related to each other needs to be articulated within an explanation.

In any case, even though *cause finding* is an important part of the process, the real objective and execution of failure analysis is to provide explanations, which becomes apparent when looking at how the practice is conducted, something which the procedure of this paper highlights. It is quite surprising then that the term "explanation" is seldom found in definitions of failure analysis. Some authors do get close by considering failure analysis as: "the science and technique of understanding how materials and products fail" (Farshad, 2011, p. 32). The hope of this paper is that definitions such as these become more standard, perhaps going a bit further and saying that the objective of failure analysis is to *explain why* failure occurred. Taking this into consideration, a possible definition could be: Failure analysis is the collection of techniques, investigative practices, and methodologies aimed toward explaining failures with the goal of achieving certain results such as corrective actions, redesign, or a better understanding of the failed system.

On a related note, the last few years have witnessed an increasing interest in computational argumentation, partly due to its similarities with intuitive human reasoning and to its relationship with explainable AI (Sklar and Azhar, 2018). As Fan and Toni (2015) show, argumentation can be considered as a process of generating explanations. The procedure presented in this paper can aid the justification of chosen explanations in terms of their underlying explanatory virtue: by linking explanatory aims to w-questions and virtues, it becomes possible to provide justification as to why such an explanation is adequate. The possibility of assigning each aim to a virtue allows to find an explanation that fits such an aim (with the possibility of querying a pre-existent database, should it be previously structured) as well as optimizing current explanations for said virtue, but it also allows to provide a justification for the adequacy of a given explanation in terms of counterfactual reasoning.[20] For instance, given an explanation with an aim to redesign (hence requiring "accuracy" as its explanatory virtue), it becomes possible to show why an accurate explanation is better (and therefore why it was chosen): it affords tracking a greater number of counterfactual dependencies, as registered by the ability to use the information presented in the explanation to productively intervene on previous designs. This can be articulated in terms of showing how the inclusion of a factor allows to answer an important w-question that the exclusion of the same factor does not afford answering (see the comparison of the hypothetical vs. the engineers' explanation in the section "Counterfactual comparison" for a practical example).

The notion of adequacy brings in the possibility of using explanations in contexts different from the ones they originated from, and it facilitates understanding into how explanations can be best suited to certain audiences and explanatory needs based on the information present within each explanation. Research on the quality of argumentations and explanations is still in its early stages, despite some interesting advancements. Wachsmuth *et al.* (2017) laid the groundwork for natural language argumentation quality; the current work expands these ideas for technical explanations (specifically engineering explanations), by adding a layer on top of quality (namely, adequacy). Moreover, conceptualizing explanations in a Woodwardian sense (i.e., as arguments that exhibit systematic patterns of counterfactual dependence, thereby affording answers to w-questions) presents an orthogonal approach that can aid the evaluation or justification of explanations that have been generated by a variety of methods.

Conclusion

This paper outlined a procedure for assessing the adequacy of failure explanations in failure analysis and applied it to the evaluation of two case studies, one from Mechanical Engineering (a broken vehicle shaft) and one from Civil Engineering (a collapse in a convention center). The procedure comprises five steps. The first two are instrumental for the rest and consist in structuring information into structural equations and identifying the set of w-questions that are relevant to the aim. The third step focuses on evaluating the quality of the explanation by exploring how well it can answer a set of relevant w-questions. The fourth step compares the explanation against competing ones or against a modified version of itself (whereby checking for improvements). The fifth concludes and gives redesign recommendations if pertinent.

The procedure offers failure analysts a tool to critically reflect on parts of their practice while providing a way to look at failure analysis as an explanatory practice. Put differently, it serves as a practical tool for evaluation and improvement of explanations, but it also helps clarify an important aspect of failure analysis.

Given the flexibility of SEM and how it has traditionally been used in areas such as psychology or sociology, a natural extension of this work is to consider socio-technical systems, which falls in line with the recent shift in perspective toward considering failure in a broader context. It would be good to see validation of the approach in these more complex cases.

Similarly, one of SEM's virtues is how it deals with error and uncertainty; this feature is not fully fleshed out in this paper (mainly because the models have already been validated), but it remains a possible avenue of application at the ground level (when building initial models), especially considering how SEM encourages the questioning and improvement of hypotheses to fit available data, compelling researchers to consider their theoretical approach as they develop their models (Rubio and Gillespie, 1995). This idea together with other considerations presents several follow-up research questions: is the procedure applicable to explanations stemming from domains other than failure analysis? Would a relative notion of relevance (where w-questions are not simply relevant or not, but relevant to different degrees) improve the usefulness of the procedure? Could this procedure be automated and integrated with other optimization strategies? Hopefully, the research presented in this paper offers a direction for thinking about these and related issues.

Acknowledgements. I would like to thank Dingmar van Eck, Erik Weber, Julie Mennes, Karim Zahidi, and four anonymous reviewers for very useful feedback on earlier versions of this paper.

Funding. Funding was provided to Kristian González Barman by the Research Foundation Flanders (FWO).

Conflict of interest. The author declares no conflicts of interest.

[20] See Sakama (2014) for a different take on using counterfactual reasoning within computational argumentation.

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press

References

Abdelhamid TS and Everett JG (2000) Identifying root causes of construction accidents. *Journal of Construction Engineering and Management* **126**, 52–60.

Affonso LOA (2006) *Machinery Failure Analysis Handbook*. Houston, TX: Gulf Publishing Company.

Andersen B and Fagerhaug T (2006) *Root Cause Analysis: Simplified Tools and Techniques*. Milwaukee: Quality Press.

Barman KG and van Eck D (2021) IBE in engineering science – the case of malfunction explanation. *European Journal for Philosophy of Science* **11**, 10.

Becker WT and Shipley RJ (eds) (2002) *ASM Handbook, Volume 11: Failure Analysis and Prevention*, 10th Edn. Materials Park, OH: ASM International.

Bell J, Snooke N and Price CJ (2007) Functional decomposition for interpretation of model-based simulation. *Advanced Engineering Informatics* **21**, 398–409.

Bhaumik S (2009) A view on the general practice in engineering failure analysis. *Journal of Failure Analysis and Prevention* **9**, 185–192.

Bloch HP and Geitner FK (2012) *Practical Machinery Management for Process Plants. Volume 2: Machinery Failure Analysis and Troubleshooting*. Amsterdam: Elsevier.

Bollen KA and Pearl J (2013) Eight myths about causality and structural equation models. In Morgan SL (ed.), *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer, pp. 301–328.

Boon M (2008) Diagrammatic models in the engineering sciences. *Foundations of Science* **13**, 127–142.

Chantler MJ, Leitch RR, Shen Q and Coghill GM (1995) A methodology for the development of model based diagnostic systems. *IEE Colloquium on Real-Time Knowledge Based Systems*, London, pp. 51–53.

Cleland JH and Jones DRH (1996) Shear failure of a road-vehicle steering shaft. *Engineering Failure Analysis* **4**, 81–88.

Delatte NJ Jr (2009) *Beyond Failure: Forensic Case Studies for Civil Engineers*. Reston, Virginia: American Society of Civil Engineers.

Dennies DP (2002) The organization of a failure investigation. *Journal of Failure Analysis and Prevention* **2**, 11–16.

de Ridder J (2006) Mechanistic artefact explanation. *Studies in History and Philosophy of Science Part A* **37**, 81–96.

Engineering News-Record (2006) Pittsburgh convention center plans to add girder seats.

Fan X and Toni F (2015) On computing explanations in argumentation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Farshad M (2011) *Plastic Pipe Systems: Failure Investigation and Diagnosis*, Vol. 407. Oxford: Elsevier.

Gabbay DM, Thagard P, Woods J and Meijers AW (2009) *Philosophy of Technology and Engineering Sciences*. Eindhoven: Elsevier.

Goel AK and Chandrasekaran B (1989) Functional representation of designs and redesign problem solving. In *IJCAI*, pp. 1388–1394.

Hall N (2007) Structural equations and causation. *Philosophical Studies* **132**, 109–136.

Halpern JY (2008) Defaults and normality in causal structures. In KR, pp. 198–208.

Halpern JY and Hitchcock C (2011) Actual causation and the art of modeling. *arXiv preprint*. arXiv:1106.2652.

Hendrick K and Benner L (1987) *Investigating Accidents with STEP*. New York: CRC Press.

Hershberger SL (2003) The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling* **10**, 35–46.

Hollnagel E (2002) Understanding accidents-from root causes to performance variability. In *Proceedings of the IEEE 7th conference on human factors and power plants*. IEEE, p. 1.

Houser M and Ritchie J (2007) Convention center collapse blamed on bolt connection. *Pittsburgh Tribune-Review*. February 22.

Jensen DC, Bello O, Hoyle C and Tumer IY (2014) Reasoning about system-level failure behavior from large sets of function-based simulations. *AI EDAM-Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **28**, 4.

Josephson JR and Josephson GS (eds) (1994) *Abductive Inference*. Cambridge: Cambridge University Press.

Katsakiori P, Sakellaropoulos G and Manatakis E (2009) Towards an evaluation of accident investigation methods in terms of their alignment with accident causation models. *Safety Science* **47**, 1007–1015.

Kleer JD and Williams BC (1987) Diagnosing multiple faults. *Artificial Intelligence* **32**, 97–130.

Laflamme L (1990) A better understanding of occupational accident genesis to improve safety in the workplace. *Journal of Occupational Accidents* **12**, 155–165.

Lehto M and Salvendy G (1991) Models of accident causation and their application: review and reappraisal. *Journal of Engineering and Technology Management* **8**, 173–205.

Leveson N (2004) A new accident model for engineering safer systems. *Safety Science* **42**, 237–270.

Li G, Gao J and Chen F (2009) A novel approach for failure modes and effects analysis based on polychromatic sets. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: aI EDAM* **23**, 119.

Martin PL (1999) *Electronic Failure Analysis Handbook: Techniques and Applications for Electronic and Electrical Packages, Components, and Assemblies*. New York: McGraw-Hill Education.

Pearl J (2009) *Causality*. Cambridge: Cambridge University Press.

Reiter R (1998) A theory of diagnosis from first principles. *Artificial Intelligence* **32**, 57–95.

Rosenblum C (2007) *Beam failure at Pittsburgh convention center fixed. Journal of the American Institute of Architects*. Available at https://www.architectmagazine.com/technology/beam-failure-at-pittsburgh-convention-center-fixed_o.

Rubio DM and Gillespie DF (1995) Problems with error in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* **2**, 367–378.

Sakama C (2014) Counterfactual reasoning in argumentation frameworks. In *COMMA*, pp. 385–396.

Saleh JH, Marais KB, Bakolas E and Cowlagi RV (2010) Highlights from the literature on accident causation and system safety: review of major ideas, recent contributions, and challenges. *Reliability Engineering & System Safety* **95**, 1105–1116.

Sklar EI and Azhar MQ (2018) Explanation through argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 277–285.

Sklet S (2004) Comparison of some selected methods for accident investigation. *Journal of Hazardous Materials* **111**, 29–37.

Stern CR and Luger GF (1997) Abduction and abstraction in diagnosis: a schema-based account. In Feltovich PJ and Hoffman RR (eds), *Expertise in Context*. Menlo Park, CA: AAAI Press, pp. 363–381.

Wachsmuth H, Naderi N, Hou Y, Bilu Y, Prabhakaran V, Thijm TA and Stein B (2017) Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, Long Papers, pp. 176–187.

Wagenaar WA and van der Schrier J (1997) The goal, and how to get there. *Safety Science* **26**, 25–33.

Wiss, Janney, Elstner Associates, Inc. (WJE). (2008). David L. Lawrence Convention Center: Investigation of the 5 February 2007 Collapse, Pittsburgh, PA, Final Report. Sports and Exhibition Authority of Pittsburgh and Allegheny County.

Woodward J (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward J and Hitchcock C (2003) Explanatory generalizations, part I: a counterfactual account. *Noûs* **37**, 1–24.

Xing L and Amari SV (2008) Fault tree analys. In Misra KB (ed.), *Handbook of Performability Engineering*. London: Springer, pp. 595–620.

Ylikoski P and Kuorikoski J (2010) Dissecting explanatory power. *Philosophical Studies* **148**, 201–219.

Kristian González Barman is a PhD candidate at the Centre for Logic and Philosophy of Science, University of Ghent, Belgium. His research focuses on explanation within the engineering sciences, causal modelling, and explainable AI.

https://doi.org/10.1017/S0890060422000099 Published online by Cambridge University Press