Two Challenges for CI Trustworthiness and How to Address Them

Kevin Baum and **Maximilian A. Köhl** and **Eva Schmidt**
Saarland University, Department of Philosophy
k.baum@mx.uni-saarland.de and mail@koehlma.de and eva.schmidt@mx.uni-saarland.de

Abstract

We argue that, to be trustworthy, Computational Intelligence (CI) has to do what it is entrusted to do for permissible reasons and to be able to give rationalizing explanations of its behavior which are accurate and graspable. We support this claim by drawing parallels with trustworthy human persons, and we show what difference this makes in a hypothetical CI hiring system. Finally, we point out two challenges for trustworthy CI and sketch a mechanism which could be used to generate sufficiently accurate as well as graspable rationalizing explanations for CI behavior.

1 Trustworthiness in Humans

For a human person to be trustworthy, she not only has to be competent at the action or decision we trust her with, but also to be appropriately motivated so to act or to decide (McLeod, 2015). To take an example from Kant (1997), the honest merchant who never cheats his customers because he worries about his reputation is someone his customers can *rely* on to be honest. He isn't trustworthy, however, for he is motivated by self-interest, not by goodwill or moral considerations. The trustworthy person is someone who has a disposition to act in a way that warrants our trust in her, for she does what we entrusted her to do for the right reasons.

Importantly, to maintain another's reasonable trust, a person has to be able to explain the motives of her actions. Imagine that you break a promise to help your friend move. This might cause her to stop trusting you—but you may avoid this result if you explain to her that you were committed to helping, but that you had to take your father to the hospital. Note that this explanation only adds to your trust*worthiness* if it is not a made-up excuse, but an *accurate* account of why you didn't do as promised, i.e., an account of your actual reasons.

2 Lessons for Trustworthy CI

These considerations make the following operationalization plausible: A CI system is fully trustworthy if and only if it (1) generally does competently what it is entrusted to do, (2) it does so for permissible reasons, (3) it is able to explain its actions[1] by reference to the reasons for which it acted, and (4) its explanations are accurate. Here, we will focus on conditions (2) to (4).

2.1 Rationalizing Explanations

What kind of explanation is sufficient to make trusting a CI system reasonable? To support our claims that we need explanations that appeal to *reasons* for which the system acted as it did (or rationalizing explanations), take the example of an automated hiring system used by a bank. Imagine that it ranks a young black woman at the bottom of the list of applicants. What would it take for her to reasonably trust that the system's decision was fair?

Clearly, an explanation in terms of a subsymbolic execution protocol[2] leading up to the decision—though it may be a true causal explanation—is

[1] Decisions are included under actions in this paper.

[2] Depending on the implementation of the subsymbolic CI system, we can obtain protocols of different forms. For instance, for a neural network we can obtain a protocol in terms of a description of all neurons and a trace of all signals.

Proceedings of the 1st Workshop on Explainable Computational Intelligence, pages 1–5, Santiago de Compostela, Spain, September 4 2017. ©2017 Association for Computational Linguistics

beside the point, for it can't help her determine whether the decision was fair/morally permissible. Rather, what is needed is a rationalizing explanation of the system's decision (Davidson, 1963), which makes the decision rationally intelligible. Rationalizing explanations appeal to the *goals* that the system pursues—sorting the applicants in order of qualification—and the *information* that it used to determine how to achieve its goals—e.g. the applicants' prior work experience. Taken together, explanations that appeal to the system's goals and information, i.e., to the reasons for which it acted, may increase its trustworthiness. In the example, if the system's explanation of its decision includes the information that the young black woman lacked the requisite work experience, she will be able to understand what motivated it and that its decision was indeed permissible.

2.2 Accurate and Permissible Explanations

Moreover, CI systems need to give *accurate* rationalizing explanations to be trustworthy. A system that gives an 'explanation' distinct from what actually drove its decision is not deserving of our trust.

Imagine that the automated hiring system excluded the young black woman not because she lacked relevant work experience, but because it is biased against people of color (Caliskan et al., 2017). If the system explains its decision by giving reasons that were causally irrelevant to the decision but make it appear permissible, there is a reason for the applicant to *reduce* her trust in it. By contrast, if an accurate explanation of what led up to its decision is provided in terms of its mistaken informational state that people of color are less qualified for the job (its bias), this will have less negative impact on the machine's trustworthiness. If we can trust that a system accurately explains its actions, we have no reason to believe that it is 'covering up' its impermissible decisions or actions.[3]

So, accurate rationalizing explanations may be given for actions that are impermissible and still make them intelligible to the persons affected. That there were certain reasons for which the system acted doesn't mean these were *good* reasons. Whether a CI system's action is permissible hinges on whether the reasons for which it was performed were permissible. To determine whether the action was permissible or not, a person then needs an accurate explanation of what motivated it.

We can distinguish *moral* (im)permissibility from *practical* (im)permissibility. A reason for which a CI system acted is morally permissible just in case it violates no moral requirements. It is practically permissible to the extent that actions motivated by it contribute to achieving what the CI system was designed to achieve.

2.3 Graspable Explanations

Further, we need CI systems to give explanations we can grasp. Assume that a system has identified a property which makes an applicant who has it the perfect employee. Its concept of the property—call it 'blagh'—is beyond our understanding. The CI system might accurately rationalize its decision by pointing out that the top applicant has *blagh*. Unfortunately, even so, we are unable to understand it, as we do not possess the concept *blagh*.[4]

2.4 Three Dimensions of Explanations

Following the insights from 2.1, 2.2 and 2.3, we can, aside from the *kind* of an explanation—merely causal vs. rationalizing—, identify the following dimensions of an explanation:

Accuracy: An explanation of an action is accurate if and only if it appeals to what actually led to the action. A rationalizing explanation is accurate if and only if it appeals to the goals and information that actually led to the action.[5]

Permissibility: An explanation is permissible if and only if the action so explained violates no moral or practical requirement. Practical requirements are given by the purposes for which a CI system is designed.[6]

[3] An accurate explanation that reveals that the system's action is impermissible may also allow us to reduce its negative impact on us. Further, accurate explanations enable the system's engineers to improve it.

[4] Cf. (Armstrong et al., 2012) and (Weinberger, 2017).

[5] As this goes to show, rationalizing explanations are *a species* of causal explanation.

[6] We have to leave to one side difficulties arising from the fact that the same action can be described in different ways. See (Anscombe, 1962). Putting these in, we get: An explanation E is permissible iff the action explained by it *under a description*

Graspability: An explanation is graspable for some person P if and only if the explanation makes use only of concepts P can grasp.

To be ideally trustworthy, a CI system needs to provide us with a rationalizing explanation which is accurate, graspable, and permissible.

3 Two Challenges for CI Trustworthiness

But how do we connect this result to the actual workings of CI systems? We use an example of a simple mechanism that sorts integers to illustrate two challenges in designing trustworthy CI systems. Let $sort$ be an arbitrary but correct sorting mechanism for lists of integers. For instance, an invocation of $sort$ on input $[3, 5, 0]$ gives output $[0, 3, 5]$.

How can we explain this output? Rationalizing explanations appeal to the *goals* that the system pursues and the *information* that it used to determine how to achieve them. A goal can be understood in terms of a specification of desired outputs, given an input. Here is a specification for *sort* using the standard model of integers: The successor of each integer within the output list, if present, is greater-equal than its predecessor and the output list contains the same integers as the input list.

3.1 The Permissibility-Accuracy Challenge

The need for goals and information raises the *Permissibility-Accuracy Challenge for CI trustworthiness*. We can give a high-level description of the goal to be set for the system: 'Choose the best applicant!'. But—unlike in the case of the *sort*—we don't know what exactly our high-level description means in terms of an input/output specification. By training the system we hope that it develops its own conception of the characteristics of a good applicant. This is what makes CI systems so powerful, but also what makes them problematic. For we cannot be sure whether a system is trained with our intended goals. Nor can we know for certain what information guides its action, so we cannot be sure whether its *actual* goals and information are permissible. To achieve trustworthiness for CI systems, then, we need to gain access to the actual goals and the information, or at least to know whether they really are permissible. As argued in Sect. 2.2, for establishing trust, it is insufficient to have *some* permissible explanation of the decision in question that doesn't reflect the actual decision-making process.

3.2 The Graspability-Accuracy Challenge

Next, it should be possible to infer from a protocol of the internal processing of the system to the information used by it to achieve its goal. In case of a classical sorting algorithm, the protocol consists of the individual symbolic steps executed.

By contrast, subsymbolic protocols of the internal processing of CI systems are not protocols of symbolic steps executed within the system, corresponding to, for instance, concepts of integer comparison or arithmetic. Rather, they provide accurate merely causal explanations, but are useless for providing rationalizing explanations. For we cannot make much sense of information presented in such monolithic form. This constitutes the *second challenge for CI trustworthiness*: We need to be able to extract or infer the information which determined the result, but also the system's *actual* goals, in terms of concepts we can grasp. As before, trustworthiness requires more than just some graspable though made-up explanation.

4 How to Meet the Challenges

To achieve CI trustworthiness, we need to tackle the two challenges: acquire rationalizing explanations which are both accurate—particularly with respect to their permissibility status—and graspable. In the following we will sketch, in a 'black box' kind of way, a mechanism that could be used to generate such explanations for CI behavior.

Formally, an accurate rationalizing explanation E consists of a set of goals G and the information which is used to determine how to achieve these goals Λ, i.e., $E := \langle G, \Lambda \rangle$. Furthermore E appeals to a model M_E which provides the concepts used within the explanation.[7] Moreover, let C denote some CI system, trained with some data D, resulting in an internal inaccessible model M and goals G. Let O_I be the output generated by C on some

matching the explanation violates no moral or practical requirement.

[7] Think of the model as including the general information about the world the system possesses, which is coached in the concepts possessed by the system.

input I. E explains O_I with respect to I if and only if it makes the output rationally intelligible.

To obtain an accurate and graspable rationalizing explanation E (or something close enough) we propose the following mechanism:

First, we need to build a *Confabulator*. Given solely an input/output series, it constructs explanatory hypotheses, i.e., candidate goals G_C and corresponding candidate explanations $E_C := \langle G_C, \Lambda_C \rangle$ appealing to a candidate model M_C which is graspable, i.e., contains only concepts we can grasp, which makes E_C graspable as well.[8] The resulting candidate explanations ignore the inaccessible actual internal model M and the actual goals G. For instance, for *sort*, we can obtain candidate explanations E_1 and E_2 based on the standard model of integers $M_{\mathbb{N}}$ and a candidate goal, which is in this case the specification S, given above:

$$E_1 : \langle S, 0 < 3; 3 < 5 \rangle \qquad E_2 : \langle S, 0 < 3; 5 > 3 \rangle$$

The Confabulator's candidate explanations have a serious shortcoming: If they are accurate at all, this is pure luck. How do we know whether $M_{\mathbb{N}}$ and the relations of *greater-than* and *less-than* play any role in the actual decision making process? Say the candidate are constructed based on a series of input lists already sorted in reverse order—if so, the mechanism could equally well have the goal of reversing lists instead of sorting them. Typically, for actual CI systems, we don't even have any specification at hand and, thus, don't know the goals. The same goes for the automated hiring system. Here, the Confabulator need to guess—or learn—the goals.

Generally, there can be multiple, mutually exclusive candidate explanations which explain the same input/output series by appeal to different goals and models, where only some of these candidate explanations are *permissible*. We call this phenomenon *explanatory underdetermination*. This can be a severe problem, e.g., there may be multiple candidate explanations of the applicant ranking, of which only some are permissible, and at the same time we are unable to verify the system directly, for we lack a specification.

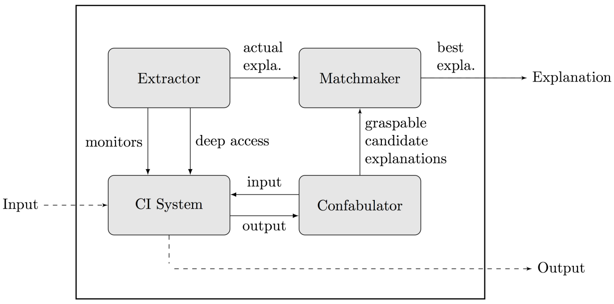

Figure 1: The configuration and interplay of a CI system with Extractor, Confabulator and Matchmaker.

This is where the second part of our proposed mechanism comes in: The *Extractor*. It extracts the Actual Explanation of the CI system's action which is probably not *graspable*, but can presumably be given, e.g., as a subsymbolic execution protocol preceding the action. The Actual Explanation may be of the wrong kind—merely causal instead of rationalizing—, but, from it, it should in principle be possible to extract or infer the reasons for which the system acted.

As a third part of the proposed mechanism, we then add a *Matchmaker*, a mechanism which matches the Extractor's ungraspable accurate explanation with the Confabulator's graspable, at best luckily accurate candidate explanations, and which then outputs a Best Explanation as *the* explanation of the CI system's decision. In so doing, the Matchmaker accords to the following principles:

(Im)permissibility Preservation: The Best Explanation is permissible if and only if the Actual Explanation is permissible.

Explanatory Equivalence: The Best Explanation is explanatory equivalent to the Actual Explanation: Any output that can be explained, on the basis of the Actual Explanation and in the light of a given input, is explainable by the Best Explanation in the light of the same input.

What allows for the possibility of the Matchmaker is that there is explanatory underdetermination. Seeing as there can be more than one explanation of the same action, we can try to move from an ungraspable actual and thus accurate explanation to a corresponding graspable explanation which preserves the permissibility status of the actual explanation. By

[8] The Confabulator may consist of human experts, of some additional system with or without access to the system under consideration, or be part of the CI system itself.

additionally requiring Explanatory Equivalence, we get accuracy—or something close enough.[9]

5 Conclusion

We have argued that for trustworthiness, CI systems have to do more than 'just their job': they have to do what they are entrusted to do for permissible reasons and to give rationalizing explanations of their behavior which are accurate and graspable. We supported this claim by drawing parallels with trustworthy human persons, and by applying our claims to a hypothetical CI hiring system. We then presented two challenges for designing trustworthy CI systems. Finally, we sketched a mechanism consisting of three components—a Confabulator, an Extractor and a Matchmaker—which could be used to generate sufficiently accurate and graspable rationalizing Best Explanations for CI behavior.[10]

This may not be the only architecture to overcome the fundamental challenges of trustworthy CI design. Difficult obstacles along the way to building our proposed mechanism are to be expected.[11] However, we believe that trying out concrete and feasible proposals for building explainable CI systems is essential to making any progress in this area at all. So, designing the three components of our proposed mechanism should be high on the research agenda of those interested in explainable CI.

References

Elizabeth Anscombe. 1962. *Intention*. Harvard University Press, Cambridge, MA.

Stuart Armstrong, Anders Sandberg, and Nick Bostrom. 2012. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines*, 22(4):299–324.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Donald Davidson. 1963. Actions, reasons, and causes. *Journal of Philosophy*, 60:685–700.

Immanuel Kant. 1997. *Groundwork of the Metaphysics of Morals*. Cambridge University Press, Cambridge.

Carolyn McLeod. 2015. Trust. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*. Stanford University.

David Weinberger. 2017. Alien knowledge: When machines justify knowledge.

[9] We have to leave out our detailed, Anscombe-inspired account of what makes a Best Explanation sufficiently accurate.

[10] Most likely, one component will fulfill all three roles at once. Note that our proposed system can use explicit *requests* for explanations as constraints on the Confabulator's output. E.g., when we want to ensure that the applicant rating CI system isn't biased against black people, we can ask the Confabulator to confabulate explanations that involve that kind of bias and then try to match it using the Matchmaker with the actual explanation extracted by the Extractor.

[11] Here's one: How can we be certain that the components of our Best Explanation Generator are trustworthy?