Minds and Machines (2026) 36:9
https://doi.org/10.1007/s11023-026-09759-y

Responsible Assessment of Beliefs Based on Computational Results: Expanding on Computational Reliabilism

Michael W. Schmidt[1] · Heinrich Blatt[1]

Received: 30 December 2024 / Accepted: 26 December 2025

Abstract

In order for advanced computational systems, such as AI systems, to be successfully integrated in liberal democracies, the people who design, use or are affected by these systems in many cases must be adequately disposed to hold the results of these systems to be true. How is such belief in these results justified, given the opaque nature of advanced computational systems and the possibility of error? The theory of "computational reliabilism" (CR) outlines how such belief can be justified and lead to a genuine advancement in human knowledge. The basic idea of CR is that the belief in the results of computational systems, despite their opacity, can be justified by their positive rate of producing true beliefs. In this paper, we show that CR needs to be expanded by focusing more on the human agents who are interacting with these systems epistemically and the consequences of the human-computer interaction. The reliability of a belief-forming process based on a computational system can only be assessed by taking into account both these agents and the ethical stakes involved. Moreover, if CR is intended to guide action, a responsible assessment of reliability must rely on an internal type of justification that is relative to the respective epistemic agent and typically necessitates an institutionalized division of epistemic labor.

Keywords AI ethics · Social epistemology · Human-computer interaction · Trustworthiness · Justification

Michael W. Schmidt and Heinrich Blatt have contributed equally to this work.

✉ Michael W. Schmidt
michael.schmidt@kit.edu

Heinrich Blatt
heinrich.blatt@kit.edu

[1] Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), P.O. Box 3640, 76021 Karlsruhe, Germany

Published online: 21 January 2026

1 Introduction

Advanced computational systems are now being used in nearly every sphere of life. However, most of these systems are largely epistemically opaque in a specific sense. The system architecture itself and its creation are so complex that no human could ever reasonably check all the various logical steps and branching points that are present in these systems. We can understand this notion more precisely with Paul Humphreys' definition of epistemic opacity:

> [A] process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. (Humphreys, 2009, 618)

Juan M. Durán and Nico Formanek identify this as the skeptical challenge that opacity poses to the epistemic usage of advanced computational systems (Durán & Formanek, 2018). This challenge, first raised by Humphreys, resulted in a lively debate, ranging from the demand for a new epistemology (Humphreys, 2009) to the rejection of the novelty of the challenge (Frigg & Reiss, 2009). If one concedes that, whether or not it is novel, there is indeed a challenge from the practical perspective of responsible epistemic agents, the epistemic opacity of advanced computational systems must be addressed.

Let us elaborate this point with the example of the joint research project KIARA, where a rescue robot is developed to assist agents of civil protection when dealing with hazards like dirty bombs (Daun et al. 2024). This project takes preexistent robot platforms and extends them to fulfill certain use cases, like measuring radiation levels while exploring rooms via teleoperation. Various hurdles might emerge during this exploration: there might be doors that need to be opened and stairs that need to be walked up, which are core challenges in robotics. To support users in teleoperation, some standard tasks like object recognition or door opening are implemented with AI-based functions. These AI components in particular incorporate many different functionalities developed by communities with an unknown number of developers. There can easily be hundreds of developers involved, with a lack of transparency as to the quality of their work as well as their validation and verification measures. So, to take up the practical challenge of epistemic opacity: How can a user of a KIARA robot, such as a police officer investigating a potential bomb laboratory, be reasonably sure that the KIARA robot is doing exactly what it is supposed to do?

With this example, it immediately becomes clear that there are many different stakeholders involved. An incomplete list of the various groups of people that should be considered in this context includes:

> *System Developers*, who integrate software (including libraries) and hardware in order to achieve a new functionality, e.g., a new computational system (including robots) that can be used by end-users;
> *End-Users*, who apply the computational system;
> *Certifiers*, who might independently ensure that the computational system is working and applied properly; and

The Public, who has an interest in the proper functioning and application of the computational system.

In order to act responsibly, the respective agents of these stakeholders cannot simply assume that the results of the computational system are correct, since an error could easily have significant moral and legal consequences. Therefore, they must properly assess the trustworthiness of the computational system. In other words, they must justify their disposition to believe or disbelieve the results of the system. Durán and Formanek's "computational reliabilism" (CR) (Durán & Formanek, 2018) tackles exactly this epistemic issue based on a given trustworthy computational system's positive rate of producing true beliefs. While this offers a good starting point, we argue that the approach of computational reliabilism is unsatisfactory in its present form and needs to be refined. In doing so, the paper proceeds at follows:

In Sect. 2, we summarize the basic idea of CR as formulated by Durán and Formanek. In Sect. 3 we show that the reliable process that CR is based upon should not be identified with a purely computational process but also includes the belief-forming process of the agents who base their beliefs on the computational process. These processes vary primarily in terms of the agents' membership in a specific stakeholder group, which entails corresponding epistemic capacities and background knowledge. In Sect. 4, we argue that there is a need to reflect on the threshold that distinguishes reliable from unreliable processes in light of the moral stakes that arise in the respective use case. In Sect. 5, we argue that for most users and persons affected by the use of a computational system as well as the general public – and to some degree also for other stakeholders – the most important resource for an adequate assessment of the reliability of the computational system is based on certification. This certification requires corresponding social institutions and practices, like the availability of construction models and industry standards. We proceed to illustrate our findings in Sect. 6 with the specific case of the rescue robots developed by the KIARA project. Section 7 provides a summarizing conclusion.

2 What is CR?

CR is inspired by classical epistemological process reliabilism (Goldman & Beddor, 2021; Goldman, 1976). In process reliabilism, a belief is justified externally if and only if it results from a reliable process that is in fact sufficiently likely to produce true beliefs in the relevant context. This external justification, as a necessary condition for knowledge, suffices to distinguish mere (lucky) true belief from an instance of knowledge. While CR is not primarily concerned with the analysis of knowledge, it builds upon process reliabilism insofar as it assumes that a trustworthy computational system needs to be a source of externally justified beliefs for a subject *S* in the process reliabilist sense:

> (CR) if *S*'s believing *p* at *t* results from *m*, then *S*'s belief in *p* at *t* is justified[;] where *S* is a cognitive agent, *p* is any truth-valued proposition related to the

> results of a computer simulation, *t* is any given time, and *m* is a reliable computational process. (Durán & Formanek, 2018, 654)

A process is understood to be reliable – this is also a technical term in CR – insofar as it has, in relevant contexts, "the tendency [...] to produce beliefs that are true rather than false" (A.I. Goldman, 1979, 95). Note that although this semi-formal explication of CR holds specifically for beliefs related to the results of computer simulations, it can be extended to all computational systems that are applied in use cases in which belief formation is crucial. Accordingly, CR has been explicitly extended to systems that are based on some forms of machine learning (Durán, 2024). However, in contrast to the primarily theoretical concept of process reliabilism, CR should, in our view, be understood primarily as a practical and action-guiding theory for real epistemic and moral agents. This is because, in addition to the external justification of beliefs based on a trustworthy computational system, where knowledge can be ascribed to an agent who forms true beliefs by means of a reliable process (i.e., a belief-forming process that is in fact truth conducive in the relevant context) irrespective of whether the agent or anyone else knows of or even believes in the reliability of that process, CR demands that agents provide explicit reasons why the process is indeed reliable (Durán & Formanek, 2018, 655f.). This requirement to provide reasons for the reliability of the process can be called the demand for internal justification of the process by a responsible agent (Durán and Formanek refer to the so-called JJ-principle (Durán & Formanek, 2018, 655)). Such a demand only makes sense if CR is conceived as a theory that aims to guide the actions of responsible epistemic agents who seek to give their trust only to trustworthy computational systems in a rational and publicly defendable way. The trustworthiness of a computational system in this sense is not transparent for actual agents – they cannot take the external point of view that is presupposed by the ascription of knowledge in process reliabilism. The need to assess the trustworthiness or reliability of the system from an internal rather than external perspective necessitates an internal justification.[1]

An important aspect of CR is the intention to bypass the requirement to confirm the epistemically oriented explainability or transparency principles that are prominent in AI ethics. The ethical requirement to provide an adequate epistemic explanation as to why a specific result of a computational system has occurred or to have full

[1] We should clarify that when we speak of "internal justification" or "external justification" this is tied to classical notions of epistemic externalism and internalism (without presupposing an externalist or internalist account). Externalism wouldn't mandate an internal justification or any other cognitive access to the facts that make a true belief an instance of knowledge (like the fact that the belief resulted from a reliable process, i.e. the external justification). Internalism mandates a cognitive access to what makes a true belief an instance of knowledge. The epistemic agent might, for example, grasp that their judgement is part of the most plausible or coherent systematization of their belief system after due reflection and thus that is in reflective equilibrium. CR is neither purely relying on external justification nor on internal justification, but is rather hybrid since it relies on both. CR is also hybrid in a different sense of external and internal justification: With regard to the epistemic agent providing an internal justification (in the previous epistemological sense) the agent needs to take into account factors that are exogenous or external to the specific computational system (like the sources of reliability that are mentioned by Durán and Formanek) and also to some degree factors that are internal to the computational system, like knowledge about the basic function of its inner working (see here also Sect. 5). We thank an anonymous reviewer for making us aware that a clarification may be needed.

transparency for all epistemically relevant steps is in many cases very demanding. If we restrict our designation of trustworthiness to systems where such a requirement of explainability or transparency is met, a relevant number of systems will not be acceptable. With CR one might justify the designation of trustworthiness without providing full explanations or transparency, and thus be more inclusive of computational systems. We are skeptical whether, from an ethical perspective, this bypassing strategy can work in all cases. In many sensitive use cases, explainability as an epistemically oriented moral principle might be a necessary requirement for the acceptability of computational systems. And transparency would be necessary in order to adhere to the principle of explainability (Martin et al. forthcoming, 2025). However, we think it is plausible that strategy of bypassing these requirements succeeds in some circumstances, where a real explanation – in the epistemic sense – would likely be too demanding in light of the potential benefits of using an opaque but de facto trustworthy computational system (see also Sect. 5). If such a modest claim is accepted, then our proposed refinement of CR in the following sections is relevant.

Durán and Formanek proceed to provide four sources of reliability which should be referred to when seeking an internal justification that renders the reliability and trustworthiness of a specific process based on a computational system sufficiently plausible (Durán & Formanek, 2018, 656-663):

1. Verification and validation methods
2. Robustness analysis
3. A history of (un)successful implementations
4. Expert knowledge

This list covers the general sources one must address when assessing the reliability of a belief-forming process based on a computational system. Interestingly, these sources are largely exogenous to the inner workings of the computational systems, and thus could to a certain degree help to bypass onerous demands for transparency and explainability in the assessment of reliability, but not completely (Alvarado, 2024) (see also Sect. 5). The sources in this list are also problematic in light of potential *adversarial attacks*, and adequately securing the reliability of the belief-forming process may require an ongoing and expanded non-standard robustness analysis that takes this into account (Pawlowski & Barman, 2025). Moreover, the list is in need of specification, especially with respect to the stakeholder groups of end-users and significantly affected persons (Sect. 5).

There are two major flaws in computational reliabilism, in our view:

1. the inadequate ascription of the property of reliability to purely computational processes,[2] and

[2] In a recent publication (Durán, 2025, 37) Durán seems to acknowledge this point: “It must be noted that CR holds that a process is broader than the algorithm qua logico-mathematical entity. It also encompasses a wider socio-techno-scientific context in which the algorithm is designed, used, and maintained.” However, he does not elaborate on the consequences of this point.

2. the lack of discussion about how to assess the threshold for the rate of truth-conduciveness of a belief-forming process based on a computational system that is necessary for that process to be regarded as reliable.

We will now elaborate on these two issues in Sects. 3 and 4, and will also offer some solutions.

3 Reliable Processes in CR Should be Understood as Relative to Epistemic Agents

Durán and Formanek ascribe the property of reliability to purely computational processes. Let's take as an example an object recognition algorithm whose results are presented on the screen of a user interface (UI) of a robot with optical sensors or an image processing medical diagnostic device. The idea here is that the algorithm is reliable in a specific context when for a sufficiently high percentage of results, e.g., that something is a door or a melanoma, the results are in fact true. To assess the "reliability" of a computational system (software and hardware) without including any human epistemic agent might make sense for fully automated machines where decisions are made without human involvement (human in or on the loop) (Grote et al., 2024). However, this does not fit with the definition of CR presented by Durán and Formanek: CR is concerned with the reliability of a belief-forming process in which it is essential that a human agent be justified in forming their beliefs based on (calibrated) trust in algorithmic results (see Sect. 2). Even if it were appropriate to ascribe beliefs or other doxastic states in the full sense to advanced computational systems like systems based on large language models (LLMs) – we are skeptical of this given the present state of these technologies – CR is still geared towards human agents who can justify that their beliefs are reliably formed and can take responsibility for those beliefs. CR is thus focused on computational systems where a human at least sometimes needs to form beliefs for proper decision making with regard to the specific use case.

Now, the belief of any human agent that interacts with a computational system, observes its results, and forms beliefs on this basis does not result from a purely computational process, as the definition of CR suggests (see Sect. 2). A crucial step on the path to beliefs or other doxastic states of a human agent is the cognitive and social process that takes place after the results of the computer system are observed (here reiterations can occur, of course). Thus in addition to computational processes, the process that leads to beliefs also includes cognitive processes. If the process that leads to beliefs is ultimately reliable, then it is the *whole* process, including the critical element of the cognitive processes that handle the observed results. And these cognitive processes will typically vary depending on the general cognitive capacities, specialization, and background knowledge of the respective epistemic agents. Therefore, the reliable processes CR is concerned with are to be understood as reliable relative to the respective epistemic agents. This agent-relative view resonates with findings from human–computer interaction (HCI) research on mental models in human–AI decision-making. As Steyvers and Kumar (2024) argue, users form mental models of AI systems that are often incomplete or distorted, which explains systematic misinterpretations such as automation bias or algorithm aversion. These insights highlight that the reliability of belief-forming processes is contingent on how epistemic agents conceptualize and internalize the functioning of the computational system.

Let us illustrate this using the earlier example of the results of recognition algorithms in robots with optical sensors or image processing medical devices. In UIs, the result is often displayed with the qualification of “confidence levels” or “confidence scores”, such as “door, 100%”. It is fairly easy to imagine that background knowledge seriously impacts the belief formation of the epistemic agents observing these results. It matters whether the user has profound knowledge of how the computational system was constructed, such as a developer of the system, or is an inexperienced user who lacks knowledge of the system’s limitations. With respect to the result "door, 100%", for example, an inexperienced user might easily form the belief that it is beyond any reasonable doubt that the robot is in fact confronted with a door, while someone familiar with the system’s limitations is likely to be more cautious. The same result may thus be involved in both a reliable and an unreliable belief-forming process for different epistemic agents, and so the property of reliability is relative to the respective epistemic agents.

Consequently, the reliability of belief-forming processes based on computational systems must be assessed with reference to the respective epistemic agents. Is the process reliable for people with profound knowledge about the construction of the computational system, or for skilled, experienced, or highly trained users, or for an average user who cannot be expected to have had much training or to possess much specialized background knowledge? A computational system’s trustworthiness in the context of CR is therefore also relative; it is trustworthy in a specific application area only when in that specific application area the corresponding human agents exhibit a reliable belief-forming process based on the results of the system.[3] This is of great importance if CR is perceived as a primarily practical and action-guiding theory (see Sect. 2).

4 Pragmatic Encroachment and CR

There is a further complication to CR that requires attention from both a practical perspective and a purely theoretical perspective. Currently, we do not have a sufficient definition for the threshold at which the tendency of a process based on a computational system to produce true beliefs is strong enough for the process to be deemed reliable. A minimal requirement would be that the rate of true beliefs must be higher

[3] We would like to stress, however, that in a different sense of the word "reliability" - already mentioned at the beginning of the section - the reliability of the computational system alone is still relevant. Whether the whole belief-forming process of an epistemic agent who bases beliefs on a computational system’s results is reliable in the sense we have defended - let’s call it "human-computer reliability" - depends to a large degree on the ability of the computational system to produce correct results that match with the facts - let’s call it "computer reliability". There are already many established ways to assess and measure computer reliability. With regards to classification, for example, accuracy, precision and recall are relevant measures.

than the rate of false beliefs. But in many instances, we would expect much higher rates of true beliefs, and would need to differentiate between false positives, false negatives, true positives, and true negatives, and corresponding limits for ascribing justified belief and knowledge. In the following, we argue that conditions for reliability vary in relation to the moral stakes in the given context, so that an agent assessing the reliability and trustworthiness of a process based on a computational system in light of CR must reflect on the specific use case and adapt the standard as well as the assessment process.

The idea that practical or moral issues can have an impact on epistemic facts like the justification of a belief or the status of knowledge is known as pragmatic encroachment. The well-known "bank cases" illustrate the issue (DeRose, 1992; Stanley, 2008): Imagine it is late on a Friday and you drive by your bank where you have planned to deposit a check, but you see long lines and would prefer not to spend much time waiting. You recall that the bank is open on Saturdays and that there are usually no lines then, so you drive straight back home, intending to come back tomorrow. It is in fact the case that the bank will be open tomorrow, but, of course, there could have been a change in opening hours that you weren't aware of. Did you know that the bank will open tomorrow? The response depends on the consequences of not depositing the check. Here we can expand the scenario and create two cases. In the first case, you have to pay a small fine because you failed to deposit the check this week, but this won't seriously harm anyone. In the second case, failing to deposit the check would for some reason have disastrous consequences; if you drive back home and the bank does not open on Saturday, the outcome would be very harmful. With these two cases in mind, one might want to ascribe knowledge to you in the first case but not in the second, where – from an epistemic perspective – you would have to check the current opening hours in order for knowledge to be ascribed. An explanation for this inclination is that the moral or practical stakes in the second case are much higher, and that the pragmatic or moral encroaches on the epistemic. There is, of course, a lively debate over whether pragmatic encroachment ought to be rejected and epistemic purism strictly adhered to (Kim 2017; Hirvelä 2023). However, if something like pragmatic encroachment is accepted (and thus some form of epistemic impurism), then there are consequences for CR.

The first consequence is that, from an external perspective, the degree of justification required for a process to be deemed reliable such that knowledge can be ascribed based on externally justified belief, would vary depending on the stakes involved. For highly sensitive decisions, such as decisions that involve medical diagnostic tools based on machine learning, the rate of true beliefs that result from the human–computer interaction must be quite high in order for reliability to be ascribed to the whole process. There might also be some differentiation in terms of true positives and false positives, and true negatives and false negatives, as well as different social groups.

The second consequence is that an epistemic agent seeking to assess the reliability of a process based on a computational system must accordingly also reflect on the stakes involved.[4] The agent's task is to provide an internal justification that shows

[4] An agent dealing with the task of identifying adequate thresholds for reliability in CR is also confronted with inductive or epistemic risk as a further complication (Biddle, 2020; Biddle & Kukla, 2017; Douglas,

that a context-dependent standard of external justification is met (or is not met) for beliefs resulting from a process based on a computational system. Interestingly, pragmatic encroachment would, in our view, not only affect the external justification of beliefs resulting from the process based on a computational system, which must be reflected in the internal justification, but would impact the internal justification as well. When the stakes are higher, it seems reasonable that the internal justification that an epistemic agent provides to themselves or others must also be more elaborate to be regarded as adequate.

5 Social Institutions for the Justification of Beliefs Based on the Results of a Computational System

Durán and Formanek discuss important sources of reliability that should be considered when assessing the reliability of a belief-forming process based on a computational system (see Sect. 2). However, their discussion overlooks a crucial element, namely the social institution of certifiers and the corresponding certificates.[5] Durán and Formanek list "expert knowledge" as one of the sources of reliability, and one could argue that the knowledge provided by certifiers is a form of expert knowledge (Durán & Formanek, 2018, 662f.). However, they focus on individual agents rather than expert institutions like certifiers, although they do acknowledge an institutional basis for individual expert knowledge. Largely following Collins and Evans (Collins & Evans, 2009), they state that:

> "[...] expertise is some sort of attribute or possession that groups of experts have and that individuals acquire through their membership of those groups." (Collins & Evans, 2009; Durán & Formanek, 2018, 662)

Here, we will focus on the institutional level and abstract from individual expertise. Expert knowledge, taken into consideration for the assessment of reliability and transmitted by expert testimony,[6] is especially important for stakeholder groups that do not possess extended background knowledge or understanding of the technical setup of a computational system. Typically, users and affected persons are stakeholders without such an extended knowledge or understanding. These stakeholders, in order to act responsibly, need to internally justify the reliability of the belief-forming process based on the computational system by reference to some trustworthy expert testimony. Since it is not an easy task for non-experts to select trustworthy experts and individual experts' expertise may vary considerably, we need a social division of epistemic labor that these stakeholders can rely upon. This is where certifiers come into play. These are institutions that are established with reference to acknowl-

2000). We thank an anonymous reviewer for bringing our attention to this issue.

[5] Recent refinements of CR acknowledge the social dimension of adequate reliability assessment (Durán, 2024). However, these refinements still omit the need for institutionalization and the specific institution of certifiers that we focus on in this section.

[6] For further complications, see (Leonard, 2023).

edged standards, are under constant scrutiny by expert groups and communities, and can assess the reliability of socio-technical systems in standardized ways. In most instances – at least in liberal democracies[7] – individual stakeholders will be justified in trusting the expert testimony of certifiers. Thus they will be in a position to build their justification of (calibrated) trust in a computational system (and of the reliability of the corresponding belief-forming process) on knowledge transmitted by these expert institutions.

There are, however, certain requirements that must be met in order for the institution of certifiers to function properly. Here, we highlight the requirement that certifiers, given that their role is already socially established, must be provided with what we call the "construction model" of the computational system. This construction model lists all the crucial design choices including a list of the main elements of the computational system, their interplay, the problems the developers envisage in specific use case scenarios, and the mitigatory measures they have taken so far. This is required because, while a higher level of transparency or explainability may be unattainable or too costly, some basic information on the inner workings of computational systems is still necessary, since it is essential for getting an idea "[...] of the sources and the nature of possible errors"[8] (Alvarado, 2024, 14).

The requirement to provide construction models is mirrored by the epistemically oriented ethical principle of transparency, and by corresponding standards for computational systems (IEEE, 2022) or corresponding regulation that demands technical documentation for general purpose AI systems (EU, 2024, Article 53, Annexes XI and XII).. As one can see now, the need for an adequate internal justification of reliability means that CR cannot bypass all issues of transparency. However, the assessment by certifiers can and in some instances certainly should use additional methods to judge the reliability of a belief-forming process based on a computational system in specific use cases: think of something like clinical trials, as they are known in the biomedical sphere (Genin & Grote, 2021; Grote et al., 2024).

In addition to the indispensability of certifiers for the responsible assessment of reliability by users and other stakeholders without expert knowledge, certifiers who check for the reliability of the epistemic human-computer interaction and the corresponding compliance with standards and regulation will, of course, be a considerable source of actual reliability as well.

[7] Liberal democracies guarantee basic rights including free speech, which is essential for public scrutiny of institutions like certifiers. Without scrutiny that can reveal misconduct, trust in these institutions is unlikely to be reasonable. We thank an anonymous reviewer for pressing us to clarify.

[8] There will be non-trivial trade-offs when trying to provide an assessment of reliability regarding information acquisition of the inner workings of the computational system and information acquisition regarding the external sources of reliability of a computational system.

6 An example: Justifying Trust in an Automatized Radiation Map within the KIARA Setting

Let us clarify our findings with a tangible example. How should trust in the automatized radiation map of a rescue robot developed in the KIARA project (see Sect. 1) be justified? In other words, how can relevant stakeholders justify the belief of the user, such as a police officer, based on the results of the computational system, as being reliable so that they can reasonably credit the agent with a beneficial knowledge gain while using the technological artifact?[9]

The first issue our findings highlight is that we must assess the reliability of the belief-forming process aided by the automatized radiation map in relation to the specific agent, in this case a police officer. The belief-forming process based on the results of the computational system is only reliable if the police officer sufficiently understands the limits of the system and is thus able to interpret the results correctly in the relevant cases. This adequate interpretation is an essential element of a reliable process. It is especially crucial in connection with the ability to correctly place trust or distrust in the system with regard to specific results in specific contexts. For example, the police officer must understand that the map is currently designed for detecting radiation from solid objects, but radiation might also be emitted from non-solid objects, such as in a contaminated laboratory. As a further example, gamma rays may overlay with alpha or beta rays, which would not be visible in the automated map in its present state. An understanding of these kinds of limitations of the computational system as well as a trained ability to place trust and distrust accordingly is essential for the end-user to reliably form beliefs and to act appropriately.

The second issue that our findings highlight is that we must reflect on the rate of true beliefs resulting from the belief-forming process based on a computational system in terms of the ethical dimension of the use case of the system. When the stakes are high from an ethical perspective, the threshold that distinguishes a reliable belief-forming process from an unreliable one will be much higher. For the KIARA robot, the stakes are very high: false beliefs, especially concerning the radiation map, could easily result in extreme danger to the public and individuals from emergency services. Therefore, in order for the process to be deemed reliable, the rate of true beliefs in typical contexts must be very high. With regard to deployment of the KIARA robot by police forces and other technical subsystems, such as automatized person detection, issues of algorithmic fairness are also highly relevant. Thus not only is the establishment of a general threshold at issue, but the rate of false and true beliefs with regard to specific societal groups is also relevant.

The third issue that our findings highlight is the importance of reflecting on what kind of knowledge we have when assessing the reliability of the process. If we are in close contact with the technical developers and with the end-user, and thus have

[9] Stakeholders can be, for example, the general public, the commanding officer of the respective emergency forces, the operating police officer themselves. These stakeholders, including the user themselves, should be interested in (internally) justifying that the user has (externally) justified beliefs because one can only adequately trust the human–machine system if the user has justified beliefs. The stakeholders need to adequately trust the system because they have stakes in its proper functioning. We thank an anonymous reviewer for pressing us to clarify.

detailed insights into the limitations of the computational system and the background knowledge of the end users, we might responsibly assess the reliability of the process by ourselves. In most cases, when we lack this knowledge, we need some trustworthy testimony regarding the technical details, the end-users' knowledge, and their interplay. In other words, we need certifiers who provide such testimony. Such certification is not just concerned with the proper technical functioning of the radiation map but must also be concerned with the interaction of the human and the machine. We especially need certification for a detailed training of the end-users that provides them with a practical understanding of the limitations of the radiation map, which are highly relevant in the envisioned use cases. For KIARA this means that the general public, persons with political responsibility or the head of operations will need to take certification into account in assessing the reliability. Typically, even the end-users themselves, in the KIARA context the robot operator, require such a certification system to adequately justify their (calibrated) trust in the system. The same is true of the technical developers, who must integrate a certification system in their development, collaborate with the potential certifiers, and provide them with the necessary information, especially the construction model of the computational system.

7 Conclusion

In specific situations, it might be reasonable to use advanced computational systems, like AI systems based on some form of machine learning, even though the epistemically oriented moral principles of explainability or transparency are not fully met. In such cases and when decision making is not fully automated – for example there is a human in or on the loop or a human simply relies on the output for further action – it is necessary to inquire whether the beliefs based on the result of these opaque systems are justified. CR is a theory that provides reliability-based criteria for such an assessment.

We have argued that CR, as currently proposed, has some serious issues and therefore needs to be revised in three specific ways:

1. CR should be concerned with the reliability of belief-forming processes that are part of epistemic human-computer interaction and not with the performances of the computational systems in isolation. The reliability of these processes thus must be assessed relative to the human agents involved, whose beliefs are essential elements of the use cases (see Sect. 3).
2. The threshold that distinguishes reliable from unreliable belief-forming processes in terms of the rate of corresponding true beliefs varies depending on differences in the ethical dimensions of the use case. When stakes are higher from an ethical viewpoint, the threshold will increase as well (see Sect. 4).
3. The adequate assessment of reliability will, in most cases, require an institutionalized division of epistemic labor. Certifiers in particular will play a crucial role (see Sect. 5).

We further clarified our findings with reference to a specific case: the automatized radiation map of a rescue robot developed within the KIARA project.

When CR is refined as we have argued, it serves as a useful theory for the justification of trust in the human-computer interaction with advanced computational systems, like AI systems, in specific cases where the demands of explainability or transparency cannot be fully satisfied, but the use of the artifacts would be very beneficial with regards to safeguarding or promoting shared values.

Acknowledgments We would like to thank two anonymous reviewers, the KIARA team, and our research group "Philosophy of Engineering, Technology Assessment, and Science" (PHILETAS) at the Institute for Technology Assessment and Systems Analysis (ITAS) at KIT for valuable comments and suggestions.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the German Federal Ministry of Education and Research (BMBF), 13N16277 (KIARA).

References

Alvarado, R. (2024). *Challenges for computational reliabilism*. [2025-08-27]https://philsci-archive.pitt.edu/23923/

Biddle, J. B. (2020). On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. *Canadian Journal of Philosophy,* 1–21. https://doi.org/10.1017/can.2020.27 (Publisher: Cambridge University Press)

Biddle, J.B., & Kukla, R. (2017). The Geography of Epistemic Risk. K.C. Elliott and T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science* (p.0). Oxford University Press. [2025-09-11] https://doi.org/10.1093/acprof:oso/9780190467715.003.0011

Collins, H., & Evans, R. (2009). *Rethinking Expertise*. Chicago, IL: University of Chicago Press. [2024-12-23] https://press.uchicago.edu/ucp/books/book/chicago/R/bo5485769.html

Daun, K., Bark, F., Tateo, D., Peters, J., Heinlein, J., Wendt, J., Heidemann, N., Kruijff-Korbayová, I., Kohlbrecher, S., Friedrich, J., Martin, D., Schmidt, M. W., Hillerbrand, R., & von Stryk, O. (2024). A Holistic Concept on AI Assistance for Robot-Supported Reconnaissance and Mitigation of Acute Radiation Hazard Situations. 2024 IEEE International Symposium on Safety Security Rescue Robotics (SSRR), 40–45. https://doi.org/10.1109/SSRR62954.2024.10770059

DeRose, K. (1992). Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research, 52*(4), 913–929. https://doi.org/10.2307/2107917

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*(4), 559–579, [2025-09-11] https://www.jstor.org/stable/188707 (Publisher: [The University of Chicago Press, Philosophy of Science Association])

Durán, J.M. (n.d.). Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms. J.M. Durán and G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives.* Synthese Library. [2024-12-16]https://philarchive.org/rec/DURMLJ

Durán, J. M. (2025). In defense of reliabilist epistemology of algorithms. *European Journal for Philosophy of Science, 15*(2), 37. https://doi.org/10.1007/s13194-025-00664-2

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

EU (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act).* [2024-12-24]http://data.europa.eu/eli/reg/2024/1689/oj/eng

Frigg, R., & Reiss, J. (2009). The Philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese*, *169*(3), 593–613, https://www.jstor.org/stable/40271311

Genin, K., & Grote, T. (2021). Randomized Controlled Trials in Medical AI: A Methodological Critique. *Philosophy of Medicine*, *2*(1), , https://doi.org/10.5195/pom.2021.27 [2024-12-29]https://philmed.pitt.edu/philmed/article/view/27 (Number: 1)

Goldman, A.I. (1979). What is Justified Belief? G. Pappas (Ed.), *Justification and Knowledge: New Studies in Epistemology* (pp. 89–104). D. Reidel.

Goldman, A., & Beddor, B. (2021). Reliabilist Epistemology. E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 eds). Metaphysics Research Lab, Stanford University. [2025-09-10]https://plato.stanford.edu/archives/sum2021/entries/reliabilism/

Goldman, A. I. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy, 73*(20), 771–791. https://doi.org/10.2307/2025679

Grote, T., Genin, K., & Sullivan, E. (2024). Reliability in machine learning. *Philosophy Compass, 19*(5), e12974. https://doi.org/10.1111/phc3.12974

Hirvelä, J. (2023). The structure of moral encroachment. *Philosophical Studies, 180*(5), 1793–1812. https://doi.org/10.1007/s11098-023-01949-z

Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese*, *169*(3), 615–626, https://www.jstor.org/stable/40271312

IEEE (2022). *IEEE Standard for Transparency of Autonomous Systems.* IEEE. [2024-12-24] https://ieeexplore.ieee.org/document/9726144 (Conference Name: IEEE Std 7001-2021)

Kim, B. (2017). Pragmatic encroachment in epistemology. *Philosophy Compass, 12*(5), e12415. https://doi.org/10.1111/phc3.12415

Leonard, N. (2023). Epistemological Problems of Testimony. E.N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 eds.). Metaphysics Research Lab, Stanford University. [2024-12-23] https://plato.stanford.edu/archives/spr2023/entriesestimony-episprob/

Martin, D., Schmidt, M. W., & Hillerbrand, R. (2025). Implementing AI ethics: The VPCIO model. AI and Ethics. https://doi.org/10.1007/s43681-025-00723-7

Martin, D., Schmidt, M. W., & Hillerbrand, R. (Forthcoming). Comparing AI Ethics and AI Regulation: Ethical Values and Principles and the Case of Well-being. In V. C. Müller, L. Dung, G. Löhr, & A. Rumana (Hrsg.), Philosophy of Artificial Intelligence: The State of the Art. SpringerNature.

Pawlowski, P., & Barman, K. G. (2025). Fortifying trust: Can computational reliabilism overcome adversarial attacks? *Philosophy & Technology, 38*(1), 21. https://doi.org/10.1007/s13347-025-00851-2

Stanley, J. (2008). *Knowledge and Practical Interests*. Oxford: Oxford University Press, Incorporated.

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science, 19*(5), 722–734. https://doi.org/10.1177/17456916231181102

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.