


RESEARCH ARTICLE

AI-Inclusivity in Healthcare: Motivating an Institutional Epistemic Trust Perspective

Kritika Maheshwari¹ , Christoph Jedan², Imke Christiaans³, Mariëlle van Gijn³,
Els Maeckelberghe⁴ and Mirjam Plantinga⁵

¹Ethics and Philosophy of Technology Section, Department of Values, Technology and Innovation, Delft University of Technology, Delft, The Netherlands; ²Ethics and Comparative Philosophy of Religion, Department of Christianity and the History of Ideas, Faculty of Religion, Culture and Society, University of Groningen, Groningen, The Netherlands; ³Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; ⁴Bioethics and Research Ethics, Faculty of Medical Sciences, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands and ⁵Department of Genetics and Data Science Center in Health, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Corresponding author: Kritika Maheshwari; Emails: k.maheshwari@tudelft.nl; Kritika136@gmail.com

Abstract

This paper motivates institutional epistemic trust as an important ethical consideration informing the responsible development and implementation of artificial intelligence (AI) technologies (or AI-inclusivity) in healthcare. Drawing on recent literature on epistemic trust and public trust in science, we start by examining the conditions under which we can have institutional epistemic trust in AI-inclusive healthcare systems and their members as providers of medical information and advice. In particular, we discuss that institutional epistemic trust in AI-inclusive healthcare depends, in part, on the reliability of AI-inclusive medical practices and programs, its knowledge and understanding among different stakeholders involved, its effect on epistemic and communicative duties and burdens on medical professionals and, finally, its interaction and alignment with the public's ethical values and interests as well as background sociopolitical conditions against which AI-inclusive healthcare systems are embedded. To assess the applicability of these conditions, we explore a recent proposal for AI-inclusivity within the Dutch Newborn Screening Program. In doing so, we illustrate the importance, scope, and potential challenges of fostering and maintaining institutional epistemic trust in a context where generating, assessing, and providing reliable and timely screening results for genetic risk is of high priority. Finally, to motivate the general relevance of our discussion and case study, we end with suggestions for strategies, interventions, and measures for AI-inclusivity in healthcare more widely.

Keywords: AI; epistemic trust; healthcare; newborn screening; technology

Introduction

Healthcare institutions worldwide are increasingly deploying artificial intelligence (AI henceforth) technologies to assist with data analysis, reduce false-positive results, and increase the accuracy of health risk predictions. Some think that advancements in machine learning and large language models hold the promise of transforming our existing healthcare systems, for instance, by boosting their potential for providing care and treatment, increasing efficiency, reducing time and costs, and easing the burden of labor for medical professionals.¹ However, realizing these benefits depends partly on the trust of those who rely on these institutions (and their members) to provide them with good quality care and support, other things being equal.

It is commonplace to think that trust from the general public and particularly from members of vulnerable groups is important at an interpersonal level, as trust in healthcare professionals fosters patients' cooperation and compliance with prevention policies and their willingness to seek treatment and care. Trust is also important at an institutional level for maintaining legitimacy of our hospitals, healthcare systems, and health organizations, for tackling significant vulnerabilities associated with seeking care, and for managing health outcomes at both population and individual levels.

Given this role and the value of the public's trust in healthcare institutions, it seems crucial to consider whether and to what extent the public can trust or continue to entrust healthcare institutions in light of what we call AI-inclusivity.² Answering this, however, requires an account of how to understand this trust in the first place. In this paper, we draw on recent literature on public trust in science and motivate *institutional epistemic trust* as an underexplored but important ethical consideration for informing the ongoing developments and implementation of AI technologies in healthcare where AI-inclusivity is deemed unavoidable, desirable, or already on the way.

As we understand it, institutional epistemic trust in AI-inclusive healthcare systems concerns the public's trust in them as providers of medical information and advice. It is complex and relational in nature and depends, in part, on the satisfaction of at least five conditions. These conditions concern the reliability of AI-inclusive medical practices and programs, the knowledge and understanding of them among different stakeholders involved, the effect on epistemic and communicative duties and burdens on medical professionals, and finally, its interaction and alignment with the public's values and interests and the background sociopolitical conditions against which AI-inclusive healthcare institutions are embedded.

To assess the applicability of these five conditions, we explore a recent proposal for making the Dutch Newborn Screening Program AI-inclusive. In its current shape, the program enjoys broad public acceptance and steady participation indicative of high institutional epistemic trust.³ Our case will help illustrate whether and to what extent implementing AI tools to screen for genetic risks hinders or fosters this trust depending on whether parents have reasons to believe that AI-inclusive screening results are reliable and responsive to their values and ideas about (the communication of) acceptable or unacceptable risks to their newborns, amongst other things.

Through our discussion, we highlight the importance, scope, and potential challenges of meeting the stated conditions of institutional epistemic trust in a context where generating, assessing, and providing timely screening information to vulnerable people are of high priority but also within other domains and areas of healthcare more widely. In doing so, we also contribute to an important lacuna in philosophical and policy work on this topic. The growing literature on the use of AI in healthcare has so far exclusively focused on some important but distinct challenges related to trust and AI.

For instance, it remains an open question whether AI itself can be considered an object of trust⁴ and whether we should reject the notion of trustworthy AI altogether.⁵ Some have questioned the impact of these technologies on patient–doctor trust relationships and whether and to what extent our existing philosophical accounts of trust can help practitioners foster clinician's trust in AI.⁶ Others have denied that trust can serve as an ethical constraint for using AI in medical decision-making.⁷ Missing from these discussions, however, is the focus on whether we can hold warranted epistemic trust in healthcare institutions that deploy AI tools in the first place.⁸

As we will see, this oversight is regrettable insofar as challenges around whether AI itself can be trusted, for instance, can affect whether the public can warrant trust in AI-inclusive healthcare institutions.⁹ Moreover, the degree of trust in these institutions may inform or help measure the degree to which patients trust medical professionals to use AI in the provision of care, offering medical advice, conducting medical research, and so on. Besides, without their trust in AI inclusive institutions, patients may not access healthcare services at all, medical professionals may be hindered in utilizing their expertise to treat those in need, and governments may be constrained in undertaking interventions in case of public health emergencies in the age of AI-inclusivity.

Our discussion is structured as follows. We start by discussing the relationship between AI-inclusivity and epistemic institutional trust. In doing so, we offer an account of the latter by spelling out its five aforementioned conditions: reliance, professional trust, communication, recipient trust, and

background conditions. Next, we explore these conditions vis-à-vis our case study. We draw some policy recommendations for informing strategies, interventions, and measures for AI-inclusivity in healthcare more widely, before concluding our paper.

AI-Inclusivity and public epistemic trust in healthcare institutions: A proposal

At the outset of our discussion, let us briefly clarify what we understand by AI-inclusivity in the present context. AI-inclusivity in healthcare is our term for referring to the inclusion or implementation of a broad range of AI technologies as tools to support and sometimes replace standard traditional tasks and methods of providing care and treatment, improving clinical decision-making, and performing various types of medical research that are otherwise typically performed by humans.¹⁰ AI-inclusivity in healthcare might be thought of, in principle, as all-encompassing if AI technologies are widely incorporated within all the distinct domains and tasks involved within the institution.

But—and this is a more realistic scenario—AI-inclusivity may remain domain-specific without healthcare institutions being wholly AI-inclusive. Domain-specific AI-inclusivity might mean that AI technologies are introduced in some domains, whilst excluding others. Among recent examples of AI-inclusivity in healthcare are the introduction of deep learning to radiology,¹¹ the introduction of natural language processing techniques to mental health screening,¹² AI-operated health chatbots for telemedicine,¹³ and assistive technologies for elderly and dementia care.¹⁴

Besides domain specificity, the introduction of AI technologies in healthcare institutions can also be task-specific. Much like domain-specific AI-inclusivity, task-specific AI-inclusivity might mean that AI technologies are introduced to perform or assist with only one specific task within a domain, such as aiding in clinical diagnosis. Or they can be used for a wide range of tasks within the same domain such as selecting therapy, making risk predictions, and stratifying patients or complex patterns in imaging data.

These aforementioned distinctions are relevant for two purposes. First, it allows us to frame our discussion in this paper within the broader category of what we might call the *ethics of AI-inclusivity*, which characterizes the inclusion or implementation of AI technologies as a morally significant transition or shift away from ordinary practices or processes. Second, it allows us to separate the public's institutional epistemic trust in domain- or task-specific AI-inclusivity within healthcare from their trust in AI-inclusive healthcare institutions as a whole. As we will see in the following discussion, these two come apart and often, for good reasons. We now turn to explaining what we mean by institutional epistemic trust.

Trust as a concept is notoriously contested, with a wide range of philosophical, social, political, and psychological concepts on offer.¹⁵ To isolate the notion we are interested in, it is helpful to start with the broad distinction between what some call practical trust and epistemic trust (Seger, unpublished ms.). Practical trust concerns trust in an entity to do certain things, such as trusting scientists to check their laboratories for contamination routinely. By contrast, when we routinely rely on the findings of scientists about, say, the risk of lung diseases from smoking because we have reasons to believe they have proper credentials and follow appropriate procedures, our trust resembles what Gürol Irzik and Faik Kurtulmus call (*basic*) epistemic trust.¹⁶

This distinction between practical and epistemic trust has clear parallels in healthcare.¹⁷ For instance, a patient might trust their doctor to have attained their medical degree with proper credentials (practical trust),¹⁸ which, in turn, may inform their judgment and belief in whether the doctor has provided them with an accurate diagnosis of their condition based on sound and reliable medical knowledge (epistemic trust). Although these examples strictly concern agential relations of trust (for instance, one's trust in scientists or doctors), the distinction between basic trust and enhanced trust can be helpfully extended to characterize our concern as public or as care or medical advice seekers with whether we can place trust in healthcare *institutions* and their members.

Although institutional practical trust tracks the competency of an institution to do or perform the functions that they are entrusted with doing, it also seems that the trust we have in healthcare institutions involves, at least in part, a kind of epistemic trust as providers of medical information and advice (such as

accurate health results) in pursuit of treatment, care, diagnosis, and so on. Although this kind of public epistemic trust takes institutions and often their members as its appropriate target or object, it allows us to situate and frame relations of agential interpersonal trust operating within healthcare settings (such as a clinicians' trust in medical AI or a patient's trust in their doctor) as being embedded in broader and complex networks within and across institutions. Besides, it also allows us to capture how distinct institutions or their members can, independently or in coordination, determine our ability to maintain or foster interpersonal trust that patients or the public at large attributes to healthcare practitioners and vice versa.

Next to characterizing the notion of trust that is relevant to our discussion, it is important to ascertain the conditions under which we can have epistemic trust in AI-inclusive healthcare institutions. For our purposes, we take as our starting point Michal Klincewicz's extension of Gürol Irzik and Faik Kurtulmus' notion of public epistemic trust in science to provide a normative formalism for institutional trust in warranted medical information and advice that healthcare institutions are tasked to provide us with. By drawing on analogies between scientific and medical practice, Michal Klincewicz's modified account of epistemic trust holds that a member of the public can place warranted basic trust in a medical professional (or medical community at large) as the provider of a medical result or advice when:

(C1) The medical professional (S) believes the medical advice or result (P) and communicates it to the member of the public (M) honestly;

(C2) M takes the fact that S believes and has communicated that P is a (strong but defeasible) reason to believe that P;

(C3) P is the output of reliable medical research or practice that S is in a position to trust;

(C4) M relies on S because she has good reasons to believe that P is the output of such medical research or practice and that S has communicated P honestly.

On this account, the special role of medical *practice* for warranted epistemic trust in the healthcare context comes to shine. Practice here refers to all the ways in which medical professional(s) and the medical profession as an institution interact with members of the public. Yet, the account is too limited in at least three important ways for our purposes.

First, it leaves too little space for disentangling the different ways various other medical profession members interact with medical AI tools in making claims about P or the distinct kinds of AI systems they might employ for generating information besides medical advice. Second, it also leaves open the specific contexts in which medical or healthcare professionals might employ AI tools for making claims about P. Finally, it also fails to capture the role and influence of distinct external background factors, such as the values and interests of patients, regulatory bodies, AI companies, as well as cultural, social, and organizational norms that might influence or undermine the public's epistemic trust in healthcare institutions.

With these points in mind, we propose modifying this otherwise helpful conceptualization of institutional epistemic trust and its application to the AI-inclusive healthcare context. According to our proposal, M (say, a screening participant) can place warranted basic institutional trust in S (the medical institution or one of its members, such as a laboratory professional, general practitioner, or physician) as a provider of P (medical result or advice based on, say, disease prediction) when:¹⁹

(C1) P is a reliable output of reliable AI-inclusive medical research and practices and/or is an extension of already existing standard non-AI medical research and practices carried out by S ("reliability condition").

- (C2) S is reasonably knowledgeable about the context of the generation of P, and S has, based on S's existing knowledge, sufficient reasons to believe that P is reliable (as specified in C1) ("professional trust condition").
- (C3) S communicates P, together with suitable information about the context of the generation of P that amounts to sufficient reasons for S to believe that P is reliable (as specified in C2) to M ("communication condition").
- (C4) M understands to a sufficient degree (i.e., commensurate with M's knowledge, education, and linguistic skills) P and the context of the generation of P (as specified in C3), so that M believes (1) that S gives correct and complete information about P and the generation of P *as S sees it* and (2) that P is reliable (as specified in C1 and C2) ("recipient trust condition").
- (C5) Neither S nor M is aware of factors in the ethical, legal, political, economic, or social sphere that facilitate and regulate the (AI-inclusive) medical research and practices with the potential to undermine the belief that P is reliable (as specified in C1 and C2) and that the AI-inclusive medical research and practices are in M's best interest ("background condition").

A schematic representation of the four conditions is as follows (Figure 1).

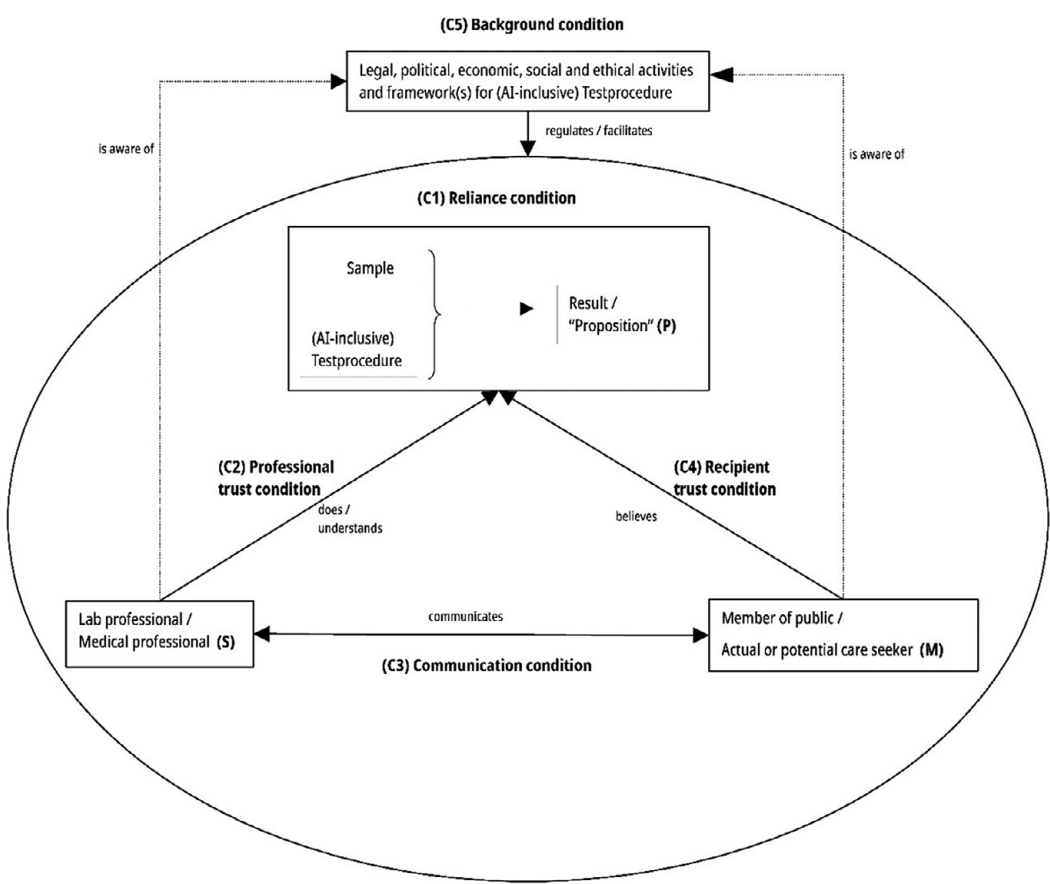


Figure 1. Five conditions of epistemic institutional trust in AI-inclusive healthcare.

As the above formalism is rather abstract, describing the conditions in some detail here is helpful to motivate our preferred formalization. Consider the first condition we call the “reliability condition.” As Michal Klincewicz correctly notes, when we trust medical experts, it is often because we think they are reliable sources of information or facts or because of some other epistemic quality or virtue of the information provided in their capacity as professionals.²⁰ In the same way, we might trust some piece of medical information or knowledge not because of our personal preferences for it but because the information or result is derived from or is a result of medical research or practice that adheres to a certain epistemically (and perhaps morally) accepted set of norms and requirements that make up for reliability.

Here, reliability tracks the objective probability of some medical result or advice (P) being true, where P is more reliable if the likelihood of it being true is sufficiently high.²¹ Reliability is essential for warranted epistemic trust insofar as unreliable medical research and practices often fail to produce truth. To see whether the reliability condition can be met, it is important then to understand to what extent AI-inclusivity impacts the reliability of P. Besides, what also matters is whether medical professionals are reasonably knowledgeable about the context of the generation of AI-aided or AI-based medical information that the public has, and that they have sufficient reasons to believe that the results are an output of reliable medical research and practice, as noted by our “professional condition.”

The third condition, namely, the communicative condition, states that members of healthcare institutions communicate P to the concerned individual(s), together with suitable information about the context of its generation that amounts to sufficient reasons for them to believe that P is reliable. An important qualifier in this condition is the word “suitable,” which characterizes the quality of the information the individual (usually the patient) should receive. It is easy for professionals in a highly specialized field to over- or underestimate the level of knowledge that care and information seekers have, resulting in miscommunication. The communication of this information thus needs to be geared to the patient’s level of understanding.

It should be noted, however, that suitable information need not be given in all cases in face-to-face communication between professionals and information seekers. Often, especially in low-risk cases and in dealing with digitally literate information seekers, the provision of general information on a website might be sufficient (see also our discussion under Section “Policy recommendations for AI-inclusivity in healthcare”). If the individual later finds out that crucial information has been withheld from them or is plainly false, then their trust risks are ruptured. This relates to what we call the recipient trust condition. Suitable communication of P requires that this communication be attuned to the communicative situation at hand, for instance, whether it is written or oral, which level of knowledge, education, and linguistic skills it assumes on the part of the individual, whether the communication is held in a timely fashion, and so on.

Concerning our final condition, it needs to be noted that AI-inclusive medical research and practice rely on an ever-growing list of ethical, legal, political, economic, and social factors and frameworks that generate medical information and advice as an output. Such factors sometimes remain in the background of (institutional) medical practice, but are significant for fostering participants’ trust. When they rise to the participants’ attention, they can do so in a negative way, acting as powerful “trust-underminers” rather than “trust-enhancers.”

For instance, the realization that the algorithm used in, say, a screening program is supplied by a multinational corporation known for selling private health data and violating privacy regulations may fuel feelings of discomfort and generalized distrust amongst the public. Conversely, the realization that an AI tool is built as open-source software by a university research hospital, without a profit interest, might enhance the participants’ trust. To illustrate further how such background conditions, along with the rest, may be understood and operationalized in specific contexts for assessing institutional epistemic trust, let us now consider a concrete case: the envisaged AI-inclusivity in the Dutch Neonatal Screening Program (Neonatal or newborn screening (NBS)).

AI-Inclusivity for newborn screening programs: a Dutch case study

NBS is a worldwide public health program aimed at the pre-symptomatic detection of rare and congenital diseases in newborns for timely interventions and improve health outcomes.²² In the Netherlands, it is conducted by the primary advisory body to the Dutch Ministry of Health, the *National Institute for Public Health and the Environment* (RIVM). Newborns are screened postpartum by a heel prick sample, collecting the blood onto a filter card, which is then analyzed in one of the five regional screening laboratories.²³ The current NBS program primarily utilizes biochemical tests, except for severe combined immunodeficiency screening involving genetic testing.²⁴

In 2020, the Dutch NBS was lauded as a successful program, as indicated by the annual monitoring figures reporting a detection rate of 1.037 per 1,000 screened children, with only two children among the 168,683 screened newborns reported as false negatives (Longaron, unpublished ms.).²⁵ The program has now expanded from 17 to 26 conditions, including autosomal recessive disorders such as cystic fibrosis. It is expected to grow more over the coming years with advancements in next-generation sequencing (NGS) techniques that allow for supplanting current biochemical tests by genetic testing and therewith detection of many more severe inherited disorders.²⁶

However, the possibility of expanding the number of diseases to screen by implementing NGS-based NBS requires the quick analysis and interpretation of hundreds of genetic variants. This can be a challenging task given the vast amount of data produced by NGS and the complexity of genomic medicine. To this end, the use of machine-learning (ML)-based AI tools can help analyze and interpret complex genomic datasets and generate and identify information on the pathogenicity of variants that were previously not possible.²⁷ One such example of an open-source and technically simple AI-based interpretation tool is the Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome Variations (CAPICE) developed by the University Medical Centre of Groningen in the Netherlands.²⁸

CAPICE is an automated system with the ability to predict whether genetic variants are potentially disease-causing (i.e., pathogenic) or whether they are non-disease-causing (i.e., benign). CAPICE uses various ML-based techniques for analyzing and recognizing patterns in the training data fed into the system. This is done using a supervised gradient-boosting tree model to generate a prediction risk score or a suggestion for classifying a genetic variant. The input and classification criteria fed as input into the automated systems allow it to classify genetic variants into pathogenic and neutral. Higher scores indicate that a variant is very likely to be pathogenic, whereas lower scores indicate a lower likelihood of pathogenicity. The predictive risk score or suggestion can be used to make decisions about follow-up testing and relevant (therapeutic) intervention in the case of a pathogenic variant suggesting a genetic diagnosis.

The use of predictive models like CAPICE is currently limited to diagnostic purposes with a higher a priori risk of a variant being pathogenic than in a screening population with low disease risk. However, its potential application to screening holds the promise of transforming existing biochemically based NBS programs into NGS-based NBS programs.²⁹ From a technical perspective, its use can aid in the interpretation of sequencing data and predict clinically relevant variants while reducing false-negative results.³⁰ From a laboratory professional's perspective, its use can help decrease and improve task efficiency by reducing the analysis time, associated costs, and the burden of manual labor that is otherwise demanded of them. Finally, from the perspective of health institutions that (plan to) introduce NGS-based NBS programs, its implementation could help achieve the goal of screening for more diseases at an early life stage and, as such, prevent or limit irreparable health damage.

Despite these potential benefits from different perspectives, it is an open question of whether AI interpretation tools like CAPICE should be used in NGS-based NBS programs. Recall that screening programs, like the ones initiated by the RIVM in the Netherlands, are *part* of broader public health programs. These programs provide medical screening information and foster screening uptake among parents who are encouraged, or sometimes even expected, to place their trust in these programs and healthcare institutions that provide for them. Parents not only rely on healthcare institutions for primary healthcare and treatment but also entrust them with providing medical information that is accurate and reliable, and timely screening results to the extent that their provision is currently a vital part of the Dutch

healthcare infrastructure and public health goals. As one report notes, “When implemented successfully, NBS programs screen >99% of newborns and are delivered with high public trust.”³¹

A big part of this effort to deliver these programs with high public trust requires and involves members of various distinct institutions like the RIVM, biochemical laboratories, bioinformatics, midwives, general practitioners, and pediatricians to acquire, process, and distribute relevant medical information, among other things. Related tasks such as reaching out to potential participants in screening, conducting biochemical tests, running data analysis using AI tools, and setting up procedures to oversee the medical information (medical results of screening) are directed toward the goal of producing medical results that can further be used to inform and carry out medical interventions to prevent or limit irreparable health damage.

In this regard, members of the general public, specifically parents interested in the early detection of congenital severe disorders for their newborns, stand in a relationship of trust with members of these aforementioned institutions. Moreover, they have an interest that these institutions not only carry out the relevant processes and tasks in screening in ways that are beneficial to us or are aligned with our goals but also serve their role and function in ways that can warrant basic epistemic trust in these institutions as providers of screening medical information. If NBS programs were to fail in this regard, then it is likely that parents have little reason to place their epistemic institutional trust in these AI-aided programs and program providers. With an eye to institutional epistemic trust, then, we can reflect and assess whether and how the proposal for AI-inclusivity in a next-generation newborn screening program can fulfill the five conditions that determine institutional epistemic trust.

Reliability condition

First, it is important to understand whether using AI interpretative tools impacts the reliability of the medical output, in this case, the screening result. Currently, CAPICE is not yet used for variant interpretation in the context of NGS-based NBS program. Evidence for the impact that AI-inclusivity would have on the reliability of the output in this specific use case is still underway. As Elaine Zaunseder *et al.* observe in their research report, “ML [machine learning] methods showed great potential in classifying NBS conditions based on screening data, their reliability has to be proven by thorough validation studies to adhere to regulatory and quality requirements before they can be integrated into NBS programs.”³²

Notwithstanding this, we can assess whether the reliability condition is met by focussing on the role CAPICE plays in this context. AI-based interpretation tools like CAPICE are an epistemic technology to the extent that they are exclusively designed to expand the epistemic capacities of medical laboratory professionals for interpreting and analyzing genetic data and producing variant pathogenicity scores. Moreover, this technology is deployed in an epistemic context of screening inquiry for processing and generating epistemic content (proposition concerning pathogenicity scores) insofar as the technology itself carries out epistemic operations such as analysis and prediction.³³

In this regard, AI-based interpretation tools in NGS-based NBS seem to serve an *epistemic* role in screening. Given this, AI-inclusivity here requires an assessment of whether the evidential norms and standards for assessing the reliance, scientific validity, and sensitivity of these tools match or align with standards typically employed for human laboratory technicians. Typically, laboratory professionals are the experts conducting research for interpreting genetic variants as pathogenic or non-pathogenic. They are entrusted by their medical peers and medical institutions at large to undertake this responsibility. The inclusion of AI-based interpretation tools for assisting in the interpretation of genetic variants thus marks a deviation from what is otherwise considered standard practice along an epistemic dimension by virtue of adding (and perhaps replacing) a new epistemic technology.³⁴

In some contexts, this deviation might be significant, but in the specific context of our proposed use case, the deviation may seem unimportant: Laboratory professionals already use interpretation tools, and the interpretation process is already partly automated, whereas the human laboratory professional retains a crucial role in monitoring and final decision-making. What would change through the inclusion

of an AI tool such as CAPICE is the nature of the automation tool that would rely on AI. In this sense, the epistemic inclusion of AI technologies such as CAPICE, within certain bounds, could be considered an *extension* of an existing practice that assigns the epistemic role of variant interpretation to human agents. This might be different in light of future technological developments, which would further affect the involvement of human laboratory professionals.

Professional trust condition

Consider, next, the professional trust condition which mostly features the laboratory professionals who will be tasked with entrusting AI-based interpretation tools while making their own assessments about the pathogenicity of genetic variants. It is, in principle, possible that the interpretation task might be entirely handled by AI-based tools while limiting the laboratory professionals' role to maintaining some level of oversight or checks and balances.³⁵ However, whether this should be the case is partly dependent on how various other ethical considerations fare in support or reasons for this limitation of laboratory professionals' involvement in the screening process.

For instance, whether laboratory professionals can be considered reasonably knowledgeable about the context of the generation of AI-based tools is complicated in light of the well-known "black-box" problem (also referred to as the opacity problem) of AI-based interpretation programs. As Raquel Dias and Ali Torkamani note, these programs are opaque technologies, such that laboratory professionals are likely to have little insight into why specific variants are predicted by the tool to be pathogenic or something else.³⁶ In the case of CAPICE, laboratory professionals often lack knowledge of the properties of a variant that has been used for considering it pathogenic, leaving it open to how these predictive AI tools make particular suggestions regarding the characteristics of genetic variants.

This knowledge gap regarding the factors underlying AI-generated specific predictions may negatively affect medical and laboratory professionals' choice of possible actions to take and evaluation of their own interpretations. Besides, we might think that even if it were, in principle, possible for laboratory professionals to get around to interpreting the AI-generated prediction scores for genetic variants, then they would nevertheless require the relevant skill, time, and resources to understand them and the potential errors contained in them, as well as finding ways to reconcile those errors. This might prove counter-productive insofar as one of the primary reasons favoring AI-inclusivity in NGS-based NBS program is that it would reduce manual labor for laboratory professionals. So, there is a trade-off to consider between the need to train staff and the potential efficiency gain and scope of the expected predictions.

Besides, how seriously the black-box problem affects the fulfillment of the professional trust condition in the use case under consideration and how that, in turn, affects the degree of warranted epistemic trust on the part of the parents may also depend on the extent to which laboratory professionals epistemically rely on the results of AI-based interpretation tools. Some level of dependence on AI-generated predictive scores may still leave open room for discretion for the professionals to decide how much time to invest, whether to review or consult their peers on the results, whether to discard negative results, and whether to treat AI-generated results as only one piece of the information that can be taken into account. Insofar as the final decision for whether a variant is pathogenic or benign ultimately rests with human laboratory professionals (and this is what is currently envisaged in the inclusion of AI in a future NGS-based NBS program), the professional trust condition remains fulfilled.

Communication condition

Consider, next, the communication condition. In our case study, the general practitioners may play a direct role in communicating the screening result based on the information or advice received by the medical advisors. The medical advisors themselves receive the screening results from laboratory professionals responsible for using AI-interpretation tools for data analysis. General practitioners may directly disseminate the result themselves, or alternatively, it may be reported indirectly through others

such as health consultants, nurses, and the printout of a laboratory report. Either way, they have the responsibility to state the result as accurately and as completely as possible, and when they directly communicate it to the parents, they have a *pro tanto* duty to report it in light of the informational needs and objectives of the parents.³⁷

The communication condition already gives rise to complex considerations, even beyond our immediate context of AI-inclusive screening programs. Physicians' conversations have been reported to need a significant amount of content judged necessary for parental understanding and sometimes contain misleading content.³⁸ Moreover, parents with low health literacy are likely to rate their primary care physicians as unable to communicate results effectively and sensitively. Experts note that there is a need to improve communication about screening results, which might become more challenging in light of the problems of opaqueness inflicted by AI-inclusivity, for instance, by delaying communication of positive screening results and thereby affecting timely interventions for treatment.

In response to the latter problem, Raquel Dias and Ali Torkamani propose that "further improvements to interpretable AI systems could not only substantially enhance the acceptability of AI predictions but also enhance the transparency of health communication between physicians and patients."³⁹ In light of our focus on epistemic trust, what would be needed, then, is a communication strategy for concrete interactions between medical professionals on the one side and parents and members of the general public on the other side to match. Although the communication of technical details might pose an undue burden on medical professionals and potentially an unnecessary source of worry on the part of parents and the general public, it also needs to be considered that public healthcare institutions need to be transparent about the inclusion of AI.

What medical professionals have to bear in mind is that the inclusion of AI as part of a new screening technique is a potential source of uneasiness on the part of the care seekers: As with every new technique and practice, the risk is that novelty is contrasted unfavorably with the familiarity of an older technique that is part of an entrusted, ongoing practice. The inclusion of AI can thus become the focus of distrust, by the mere fact of it being an innovation, with patients and the general public lacking a clear idea regarding the functioning of AI systems or having heard stories of AI going rogue. It should also be considered, however, that familiarity and acceptance are a shifting frontier: Medical institutions and professionals likely have to make a stronger effort at communication in the early stages of the adoption of an AI-inclusive practice compared to later stages.

Recipient trust condition

The fourth condition focuses on the healthcare or information seeker's ability to understand, to a sufficient degree, the screening result of an AI-inclusive procedure and the context in which it was generated, including the correctness of the assessment given by the medical professional(s). In the specific context of newborn screening, Beth Tarini has noted that "[C]oncerns about the potential for parents to misunderstand newborn screening are well founded. It is widely accepted that parents are woefully undereducated about newborn screening... The causes of poor parental understanding are multifactorial. First, testing is mandatory, but education is not... Second, parents are emotionally and physically exhausted after the birth of their child, making it difficult for them to learn and retain information about newborn screening."⁴⁰

Although this analysis might not generalize to Dutch NBS programs, where participation in the screening is optional, and information about the screening is already given during pregnancy, it highlights that AI-inclusivity may exacerbate the already "epistemic vulnerability" of young parents at a difficult time. It is important, thus, to realize that although patients, parents, and members of the general public cannot be expected to attain an in-depth understanding of AI-interpretative tools and their functioning, it is in their interest to understand at least the principles undergirding AI-inclusive healthcare. Besides, in light of the previous condition, they are owed communication that is commensurate with their level of understanding. Of course, we also see here a task for governments to make

school curricula at least “AI-aware” and for medical professionals to inform the general public through lectures, personal consultation programs, or other educational events.

Background condition

Consider, then, the final condition concerning the background factors in the ethical, legal, political, economic, or social sphere that facilitate and regulate the AI-inclusive medical research and practices. These factors have a potential to undermine the belief that screening results are reliable and that AI-inclusive medical research and practices are in the parents’ (and their newborn’s) best interest. An example of how background conditions may undermine epistemic trust is when parents find out that the algorithm used in an AI-inclusive screening procedure is not only provided by a large company with an undeniable profit interest but also provided by the company that stores and shares data outside the regulatory framework of the European Union General Data Protection Regulation (EU GDPR).

Such pieces of information might act as powerful “trust underminers,” underscoring the need to have proper safeguards also at the contextual level of AI-inclusive medical research and practices to maintain epistemic trust. Additionally, violations of the background condition will usually affect other trust conditions. To continue with the above example, if the provider of the crucial algorithm were to restrict, for reasons of competitive advantage, the provision of information about the algorithm to researchers and practitioners, this would clearly affect all the previous conditions. It appears that the involvement of AI increases the burden on medical institutions and their members to establish and communicate also the trustworthiness of the context in which they operate.

Conversely, the fact that in our case study, an algorithm was developed by a public institution such as a university hospital on a not-for-profit basis and that the algorithm and code base is open-source software (and thus open to scrutiny by stakeholders) could provide an important contribution to the fulfillment of this condition. However, this might come at the cost of compromising accuracy or empowering AI models with more capacities. One reason for thinking so is that for-profit industries may have more resources to develop powerful AI models, which public organizations often lack. Given this, they are better positioned for developing more accurate and reliable AI interpretation tools, thereby leading us back to the issues of ensuring their trustworthiness previously outlined.

Policy recommendations for AI-inclusivity in healthcare

We are now in a position to offer seven policy recommendations for concrete strategies, interventions, and measures aimed at sustaining or at least preventing the erosion of institutional epistemic trust on the part of healthcare seekers and the general public. Our hope is that these recommendations contribute to facilitating and realizing the development and implementation of AI technologies in healthcare institutions more widely, especially in light of various social, political, and technical trends toward AI-inclusivity in healthcare that we are currently witnessing. Although by no means exhaustive, the recommendations offered below closely align with and reflect the conditions for institutional epistemic trust in the order we have discussed above.

- (i) Medical institutions have strong reasons to adopt a “conservative” approach in the introduction of AI-inclusive medical practices (as opposed to research), focusing on the extension of existing (non-AI) medical practice. A good strategy is the enhancement of very specific (and limited) tasks by AI-inclusive tools, under close supervision by human expertise, so that human professionals retain control over, and end responsibility for, the specific task AI is used for (relates to reliability condition).
- (ii) Medical professionals should receive continuous training on the workings of the different AI tools they use. Whenever possible, they must shoulder the burden of ascertaining the functioning of AI tools and the adequate (direct or indirect) communication of workings and results to healthcare seekers and the public (relates to the professional trust condition).

- (iii) Healthcare institutions need to be transparent about the use of AI technology. Initiatives such as the development of an “algorithm register” in the Netherlands, aimed at describing the use of AI by governmental bodies, is a good example of the promotion of transparency (relates to the communication condition).⁴¹
- (iv) Medical professionals have a *pro tanto* duty to communicate the medical results but also, when applicable, explain the working and reliability of AI-inclusive medical practice. Although this duty could be discharged, at an initial level and where low-risk uses of AI (such as an envisaged inclusion of AI in the Dutch NBS) are concerned, though, for instance, an accessible-formulated website, to which patients, families, and general public are directed, a human interlocutor should be available to answer more targeted questions. In order to discharge this duty, medical staff fulfilling this public-facing role need to receive enhanced training in communication so that they can respond to the emotional and epistemic needs of their interlocutors (related to the communication condition). For instance, a website in an easily accessible language could serve as a way to initially fulfill the duty to offer suitable information, at least for digitally literate information seekers. This could be accompanied by the offer in case of persisting worries or questions to see a medical professional for further information.
- (v) With AI-inclusive medical practice on the rise for a number of reasons (personnel shortages, efficiency gains, balancing the effects of aging populations, etc.), there is a need to enhance “AI literacy” on the part of the general public to avoid a blanket distrust of medical innovation. This may be achieved by organising knowledge dissemination programs for explaining the principles and use of medical AI as part of school curricula, public lectures, health information drives, consultation sessions and the like organized by medical faculties and university hospitals (relates to the recipient trust condition). For such activities, enhanced training in communication on the part of medical staff and public healthcare workers might prove important (relates to the communication condition).
- (vi) The primary drivers of innovation regarding AI-inclusive research and practice should be organizations that explicitly subscribe to generating and sharing knowledge aimed primarily at generating and distributing public health goods, not commercial profits (relates to the background condition).
- (vii) Medical institutions should choose from the available option standards that ensure maximal accountability and transparency: Algorithms and programs should be developed in compliance with open-source standards; data obtained by AI-inclusive research and practice should be shared on the basis of academic standards of reciprocity (relates to the background condition).

Conclusions

In this paper, we have argued that the responsible development and implementation of AI rely on maintaining high levels of trust. We have interpreted this trust as epistemic and have proposed a framework of five conditions that need to be met. Healthcare institutions and organizations must focus on providing the right conditions that ground or allow individuals to foster warranted trust and acknowledge that doing so faces several challenges that need to be addressed. We have argued, first, that AI-inclusivity in screening practices may violate the condition of reliability of medical results by virtue of failing to be an appropriate extension of existing healthcare research and practice for epistemic reasons. Second, AI-inclusivity may challenge trust by undermining a medical professional’s trust because of the problem of opacity facing AI tools. Third, AI inclusivity has the potential to generate new communicative duties for (laboratory) professionals. Fourth, we have argued that AI-inclusivity places high burdens on care and information seekers’ digital and health literacy. Lastly, we have argued for the importance of contextual factors surrounding and structuring the introduction of AI-inclusive research and practice.

Acknowledgements. We are indebted to the members of the ELSA AI Lab and attendees of the ELSI research meetings for their valuable feedback on earlier versions of this paper. In addition, we would also like to thank Michael Gregory, Andreas Gammon, and Lavinia Marin for helpful discussions on the topic.

Funding statement. This publication is part of the project ELSA AI Lab Northern Netherlands (ELSA-NN) (with project number NWA.1332.20.006) of the Dutch Research Agenda (NWA) AI Synergy program “Artificial Intelligence: Human-centred AI for an inclusive society—towards an ecosystem of trust,” which is (partly) financed by the Dutch Research Council (NWO).

Competing interest. The authors declare none.

Notes

1. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education* 2023;23:689. doi:[10.1186/s12909-023-04698-z](https://doi.org/10.1186/s12909-023-04698-z); Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal* 2021;8(2):e188–94. doi:[10.7861/fhj.2021-0095](https://doi.org/10.7861/fhj.2021-0095).
2. We say more about this in Section “AI-inclusivity and public epistemic trust in healthcare institutions: a proposal.”
3. Van Der Burg S, Verweij M. Maintaining trust in newborn screening: Compliance and informed consent in the Netherlands. *Hastings Center Report* 2012;42(5):41–7. doi:[10.1002/hast.66](https://doi.org/10.1002/hast.66).
4. Al P. (E)-Trust and its function: Why we shouldn’t apply trust and trustworthiness to human–AI relations. *Journal of Applied Philosophy* 2023;40(1):95–108. doi:[10.1111/japp.12613](https://doi.org/10.1111/japp.12613); Ferrario A, Loi M, Viganò E. Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics* 2021;47(6):437–8. doi:[10.1136/medethics-2020-106922](https://doi.org/10.1136/medethics-2020-106922); Ryan M. In AI we trust: Ethics, artificial intelligence, and reliability. *Science & Engineering Ethics* 2020;26(5):2749–67. doi:[10.1007/s11948-020-00228-y](https://doi.org/10.1007/s11948-020-00228-y).
5. Braun M, Bleher H, Hummel P. A leap of faith: Is there a formula for “trustworthy” AI? *Hastings Center Report* 2021;51(3):17–22.
6. Viehoff J. Making trust safe for AI? Non-agential trust as a conceptual engineering problem. *Philosophy & Technology* 2023;36(4):64. doi:[10.1007/s13347-023-00664-1](https://doi.org/10.1007/s13347-023-00664-1); Nickel PJ. Trust in medical artificial intelligence: A discretionary account. *Ethics & Information Technology* 2022;24(1):7. doi:[10.1007/s10676-022-09630-5](https://doi.org/10.1007/s10676-022-09630-5).
7. Hatherley JJ. Limits of trust in medical AI. *Journal of Medical Ethics* 2020;46(7):478–81. doi:[10.1136/medethics-2019-105935](https://doi.org/10.1136/medethics-2019-105935).
8. To the best of our knowledge, only one contribution (which we will discuss in Section “AI-inclusivity and public epistemic trust in healthcare institutions: a proposal”) assesses the role of AI in mediating trust in medicine as an institution. See Klinecicz, M. Institutional trust in medicine in the age of artificial intelligence. In: Collins D, Alfano M, Jovanovic I, eds. *The Moral Psychology of Trust*. Lanham, MD: Rowman and Littlefield/Lexington Books; 2023:259–74.
9. Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: The challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association* 2021;28(4):890–4. doi:[10.1093/jamia/ocaa268](https://doi.org/10.1093/jamia/ocaa268).
10. Understood this way, AI-inclusivity differs from concerns around inclusivity *within* AI, which closely relates to issues that arise in the context of the so-called value alignment problem in AI. See Gabriel I. Artificial intelligence, values, and alignment. *Minds & Machines* 2020;30(3):411–37. doi:[10.1007/s11023-020-09539-2](https://doi.org/10.1007/s11023-020-09539-2).
11. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, et al. Deep learning in radiology. *Academic Radiology* 2018;25(11):1472–80. doi:[10.1016/j.acra.2018.02.018](https://doi.org/10.1016/j.acra.2018.02.018).
12. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: A systematic review and research framework. *Translation Psychiatry* 2023;13(1):309. doi:[10.1038/s41398-023-02592-2](https://doi.org/10.1038/s41398-023-02592-2).
13. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D’Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital Health* 2023;9:1–11. doi:[10.1177/20552076231183542](https://doi.org/10.1177/20552076231183542).

14. Vollmer Dahlke D, Ory MG. Emerging issues of intelligent assistive technology use among people with dementia and their caregivers: A U.S. perspective. *Frontiers in Public Health* 2020;8:191. doi:10.3389/fpubh.2020.00191.
15. Baier AC. Trust and antitrust. *Ethics* 1986;96(2):231–60; Hawley K. Trust, distrust and commitment. *Noûs* 2014;48(1):1–20. doi:10.1111/nous.12000; Simpson TW. What is trust? *Pacific Philosophical Quarterly* 2012;93(4):550–69. doi:10.1111/j.1468-0114.2012.01438.x; Tallant J. You can trust the ladder, but you shouldn't. *Theoria* 2019;85(2):102–18. doi:10.1111/theo.12177.
16. Irzik G, Kurtulmus F. What is epistemic public trust in science? *British Journal for the Philosophy of Science* 2019;70(4):1145–66. doi:10.1093/bjps/axy007. These authors contrast basic epistemic trust with what they call *enhanced* epistemic trust. The latter accounts for the idea that in addition to trusting scientists as providers of information, we also often care about whether they respond to and align with our values and make well-informed decisions for us. For alternative ways of spelling out epistemic trust in science, see Wilholt T. Epistemic trust in science. *British Journal for the Philosophy of Science* 2013;64:233–53.
17. Notably, the relationship between public trust in healthcare institutions is under-explored. Gille F, Smith S, Mays N. Why public trust in health care systems matters and deserves greater research attention. *Journal of Health Services Research & Policy* 2015;20(1):62–4. doi:10.1177/1355819614543161.
18. For a discussion of conceptions of “practical” public trust in healthcare systems, see Gille F, Smith S, Mays N. Towards a broader conceptualisation of ‘public trust’ in the health care system. *Social Theory Health*. 2017;15(1):25–43. doi:10.1057/s41285-016-0017-y; Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine* 2020;1(2):100001. doi:10.1016/j.ibmed.2020.100001; Goold SD. Trust and the ethics of health care institutions. *Hastings Center Report* 2001;31(6):26. doi:10.2307/3527779.
19. Note that our modification involves both reordering of the conditions and inclusion of a new one. See [note 8](#), Klinecicz 2023.
20. See [note 8](#), Klinecicz 2023.
21. We acknowledge that there are disagreements about how to assess reliability, and it may differ contextually. For our purposes, we are appealing to a basic and commonly accepted notion of simplicity.
22. Bick D, Ahmed A, Deen D, Ferlini A, Garnier N, Kasperaviciute D, et al. Newborn screening by genomic sequencing: Opportunities and challenges. *International Journal of Neonatal Screening* 2022;8(3):40. doi:10.3390/ijns8030040.
23. Jansen ME, Klein AW, Buitenhuis EC, Rodenburg W, Cornel MC. Expanded neonatal bloodspot screening programmes: An evaluation framework to discuss new conditions with stakeholders. *Frontiers in Pediatrics* 2021;9:635353. doi:10.3389/fped.2021.635353.
24. Blom M, Bredius R, Weijman G, Dekkers E, Kemper E, Van Den Akker-van Marle M, et al. Introducing newborn screening for severe combined immunodeficiency (SCID) in the Dutch neonatal screening program. *International Journal of Neonatal Screening* 2018;4(4):40. doi:10.3390/ijns4040040.
25. The newborn blood spot screening in the Netherlands - *Monitor* [Internet]; 2020; available at <http://www.neorah.nl/>.
26. Veldman A, Kiewiet MBG, Heiner-Fokkema MR, Nelen MR, Sinke RJ, Sikkema-Raddatz B, et al. Towards next-generation sequencing (NGS)-based newborn screening: A technical study to prepare for the challenges ahead. *International Journal of Neonatal Screening* 2022;8(1):17. doi:10.3390/ijns8010017.
27. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine* 2019;11(1):70. doi:10.1186/s13073-019-0689-8.
28. Li S, Van Der Velde KJ, De Ridder D, Van Dijk ADJ, Soudis D, Zwerwer LR, et al. CAPICE: A computational method for consequence-agnostic pathogenicity interpretation of clinical exome variations. *Genome Medicine* 2020;12(1):75. doi:10.1186/s13073-020-00775-w.
29. O'Brien TD, Campbell NE, Potter AB, Letaw JH, Kulkarni A, Richards CS. Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the identification of new positive cases and reduces

- analysis time in a clinical diagnostic laboratory. *Genetics in Medicine* 2022;24(1):192–200. doi:10.1016/j.gim.2021.09.007.
30. Zaunseder E, Haupt S, Mütze U, Garbade SF, Kölker S, Heuveline V. Opportunities and challenges in machine learning-based newborn screening—A systematic literature review. *JIMD Reports* 2022;63(3):250–61. doi:10.1002/jimd.12285.
 31. Vears DF, Savulescu J, Christodoulou J, Wall M, Newson AJ. Are we ready for whole population genomic sequencing of asymptomatic newborns? *Pharmacogenomics and Personalized Medicine* 2023;16:681.
 32. See note 30, Zaunseder et al. 2022, at 259.
 33. For a discussion of what makes AI an epistemic technology, see Alvarado R. What kind of trust does AI deserve, if any? *AI Ethics* 2023;3(4):1169–83. doi:10.1007/s43681-022-00224-x.
 34. See note 33, Alvarado 2023.
 35. Note that for contexts outside of NGS-based NBS screening, the professional trust condition will refer to other involved professionals.
 36. See note 27, Dias, Torkamani 2019. Note that the type of opacity in question is fundamental opacity, meaning that the decision procedures of machine learning algorithms, which work by a mathematical process of iterative statistical optimization, resist interpretation in terms comprehensible to any or most humans.
 37. Fay M. Negligence and the communication of neonatal genetic information to parents. *Medical Law Review* 2012;20(4):604–30. doi:10.1093/medlaw/fws024.
 38. Farrell MH, La Pean A, Ladouceur L. Content of communication by pediatric residents after newborn genetic screening. *Pediatrics* 2005;116(6):1492–8. doi:10.1542/peds.2004-2611; Tarini BA, Goldenberg AJ. Ethical issues with newborn screening in the genomics era. *Annual Review of Genomic Human Genetics* 2012;13(1):381–93. doi:10.1146/annurev-genom-090711-163741.
 39. See note 27, Dias, Torkamani 2019, at 9.
 40. See note 38, Tarini, Goldenberg 2012, at 385.
 41. See <https://algoritmes.overheid.nl/nl>.