BJHS Themes (2023), **8**, 127–141
doi:10.1017/bjt.2023.5

RESEARCH ARTICLE

Errors and fallibility in radiology: X-ray readings and expert radiologists, 1947–1960

Flora Lysen

Department of Society Studies, Maastricht University, the Netherlands
Email: f.lysen@maastrichtuniversity.nl

Abstract

This article traces the historical emergence of a new understanding of radiologists as fallible expert observers from the late 1940s, a conception that was shaped by new technologies and techniques, but also prepared the ground for promises of automation and artificial intelligence in the field of medical imaging. Reports of radiologists' unreliable performance prompted investigations in many countries into 'observer variability' and 'observer error'. Towards the end of the 1950s, scientists could conceive of radiologists as imperfect medical decision makers, while they concurrently developed a new model for 'logical analysis' of the diagnostic process that would limit errors. As well as technological solutions to flawed X-ray readers, researchers proposed 'double-reading' practices (a second independent reading) as a way to mitigate the 'human factor'. Yet these ideas did not find widespread resonance due to concerns about feasibility and debates about radiological expertise, and also because of a discrepancy between experimental models and real-world practices. A genealogy of the fallible trained observer helps us understand persistent worries about – and solutions to – radiologists' 'error problem' and contributes to a better understanding of current discourses on AI in medical imaging.

'AI beats radiologists' and 'Algorithms outperform doctors' are examples of the excited headlines that promise that artificial intelligence is drastically changing healthcare. Today, when advocates of deep learning aim to illustrate the powers of AI for image and pattern recognition in big data sets, advances in medical imaging are often a key example. These developments have also turned the field of radiology into a central site to study configurations of humans and machines in the development of new norms and forms of automation and artificial intelligence. As I will show, radiologists have been at the forefront of research into technologies and computational techniques prospected to fundamentally alter the work of trained, highly skilled medical professionals since the 1950s. Such promises of radical transformation through technological innovations are long-standing, but also equivocal. Evidence from experimental, laboratory settings shows that AI-supported image recognition in X-rays may help to detect a suspicious area in a scan more quickly and can classify abnormalities with more precision than experts.[1] However, as yet there

1 Xiaoxuan Liu *et al.*, 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis', *Lancet Digital Health* (2019) 1(6), pp. 271–97.

is no conclusive proof of their efficacy in clinical practice.[2] The extent to which radiologists will actually benefit from integrating new image recognition software into their everyday work routines remains unclear. Contemporary studies of algorithms trained on databases of chest X-ray images found that such models may exacerbate existing gender and racial biases and lead to more disparities in care.[3] Moreover, recent attempts to use deep-learning algorithms to detect COVID-19 in chest X-rays have failed to deliver impactful results.[4] According to some authors this technological overpromise has caused a 'credibility crisis' for machine learning in medicine.[5]

Despite these uncertainties about future benefits, both popular news reports and professional discussions about AI and medical imaging abound with bombastic metaphors expressing the extraordinary promise of big-data analytics, machine learning and deep learning to improve healthcare. The notion of 'augmented medicine', for example, envisions data technologies as extensions of human medical professionals to improve clinical practice.[6] 'Deep medicine' conjures the dream of a 'total archive' of health data to detect valuable patterns: correlations beyond the capacities of human perception and cognition.[7] Artificial intelligence is said to give rise to the 'robot radiologist', a figure that represents (anxieties about) fully automated medical image recognition, replacing human radiologists in the foreseeable future.[8] The trope of the 'centaur radiologist' conjures a synergistic image of human plus computer harmoniously combining human skilfulness with the newest AI technologies.[9] Collectively, these currently pervasive imaginaries bind together intersecting promises. In radiology, AI is thought to be able to make inferences about medical data that go beyond human interpretations, speed up routine tasks and free up time for meaningful patient contact, alleviate radiologists from tedious and repetitive work, help radiologists cope with a data deluge of records and medical images, and make care more affordable by being more cost-effective. Perhaps most importantly, however, artificial intelligence is envisioned as a technological aid to radiologists, who need this assistance because their capacities for perceiving and interpreting images are imperfect – human professionals inevitably make mistakes. AI's central promise in medicine is to reduce, or even eliminate, human error.

This article traces the historical emergence of the idea of the fallible radiologist, a conception shaped by attention to new techniques and technologies that also prepared the ground for promises of automation and artificial intelligence in medicine. My genealogical account foregrounds this oscillation or interplay between ideas about

2 Laure Wynants *et al.*, 'Prediction models for diagnosis and prognosis of Covid-19: systematic review and critical appraisal', *British Medical Journal* (2020) 7(369), m1328.

3 Seyyed-Kalantari *et al.*, 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations', *Nature Medicine* (2021) 27(12), pp. 2176–82.

4 Michael Roberts *et al.*, 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for Covid-19 using chest radiographs and CT scans', *Nature Machine Intelligence* (2021) 3(3), pp. 199–217.

5 Casey Ross, 'Machine learning is booming in medicine. It's also facing a credibility crisis', *STAT*, 2 June 2021, at www.statnews.com/2021/06/02/machine-learning-ai-methodology-research-flaws (accessed 21 June 2022).

6 Giovanni Briganti and Olivier Le Moine, 'Artificial intelligence in medicine: today and tomorrow', *Frontiers in Medicine* (2020) 7(27), pp. 1–6.

7 On the rhetoric of completeness and the total archive connected to 'deep medicine' see Lukas Engelmann, 'Into the deep: AI and total pathology', *Science as Culture* (2020) 29(4), pp. 625–9.

8 Sara Reardon, 'Rise of robot radiologists', *Nature* (2019) 576(7787), pp. 54–8.

9 Dave Pearson, 'RSNA 2016: radiologists must become as centaurs joined at the midsection to AI' (30 November 2016), at www.radiologybusiness.com/topics/health-it/enterprise-imaging/imaging-informatics/rsna-2016-radiologists-must-become-centaurs (accessed 21 June 2022). See earlier tropes of AI–human centaurs in research on chess playing, for example Kevin Kelly, 'The centaur revolution' (16 April 2018), at www.iftf.org/future-now/article-detail/the-centaur-revolution (accessed 21 June 2022).

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

human capacities and the potential of technological aids: histories of artificial intelligence, I emphasize, are also histories of imaginaries of human (in)competences. New conceptions of flawed radiologists created space for solutions by new computational techniques and technologies. Focusing on claims about improving the reading of X-rays, I analyse professional discourses that were initiated in the US but also occurred in dozens of other countries between the mid-1940s and the early 1960s. My genealogy of the flawed expert radiologist begins with an analysis of reports on 'observer variability' or 'observer error' in mass screening campaigns for tuberculosis and pneumoconiosis (coalminer's 'black lung disease'). These reports alarmed researchers with high levels of disagreement and inconsistency in the assessments of X-ray images, and also stimulated a search for solutions – often, but not always, mixing technological aids with human expertise. I show first how these debates led to the articulation of the radiologist as a fallible observer and then how, in the 1950s and 1960s, statistical and logical analyses of diagnosis, and the design of many technological 'aids' and techniques to assist diagnosis, reframed the fallible observer as a suboptimal decision maker. The final section considers one particular suggested solution. Numerous reports published in the 1950s provided evidence that a 'double reading' or 'dual reading' of X-rays by the same or another radiologist could diminish the number of overlooked anomalies. To understand why this human solution to the 'human factor' of error in radiology was hardly implemented, while promises of technological aid persisted, I will argue that we need to consider social and professional status in this field of changing labour, and expertise under question.

'The hard fact of their own unreliability': radiologists discover observer variability

'Why not X-ray before marriage?' an American health officer wondered in a medical journal in 1950, considering premarital chest X-ray films as a routine procedure to guard against the spread of tuberculosis.[10] With cheaper, mobile diagnostic imaging facilities, X-raying had become ubiquitous. By the mid-1940s, mass miniature radiography services that produced small-sized photofluorograms had emerged around the world to serve mass health surveys (predominantly to catch cases of tuberculosis), including pre-employment screenings (for medical personnel, food handlers and schoolteachers, for example) and population-wide medical examinations.[11] Public institutions were amassing immense numbers of images of citizens' chests. The Veterans Administration, for example, kept a minimum of two on file for each US soldier – one on entry and one on leaving. Giant databases of images and medical records were growing at impressive speed.

With an eye to the rapidly rising number of X-ray images taken during the Second World War, the administration of veterans' affairs sought to determine which type of X-ray technology was diagnostically the most efficient. This was particularly important because diagnostic decisions were usually made on the basis of a single miniature X-ray. While patients in clinics received multiple tests, in mass screening programmes diagnosis depended on very limited observations in a population of subjects, many of whom were symptom-free. Did a routine small-sized photofluorogram, stereophotofluorograms, a roentgenogram negative on paper or a celluloid film perform best?[12] Diagnostic efficiency was defined as those images that produced the lowest amount of under-reading (misses) of X-rays with evidence of tuberculosis as well as the

10 C.V. Craster, 'Why not X-ray before marriage?', *Journal of the Medical Society of New Jersey* (1950) 8(47), p. 395.

11 Joseph D. Wassersug, 'Common pitfalls in the X-ray diagnosis of tuberculosis', *New England Journal of Medicine* (1951) 245(16), pp. 598–600.

12 Carl C. Birkelo *et al.*, 'Tuberculosis case finding: a comparison of the effectiveness of various roentgenographic and photofluorographic methods', *Journal of the American Medical Association* (1947) 133(6), pp. 359–66.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

lowest amount of over-reading (finding a false positive).[13] A team of radiologists, lung experts and a 'biostatistician' (a specialty that rose to prominence in the 1930s) set out to compare the various machines and procedures. The answers to their questions were published in the *Journal of the American Medical Association* in 1947 and left the scientific community bewildered.

Not one X-ray method, not even the relatively expensive celluloid X-ray image routinely used in hospital clinics, appeared to allow better performance over others in finding tuberculosis cases. A far more significant result emerged: when hundreds of image evaluations made by five readers were compared, these experts appeared to have a very high degree of disagreement and inconsistency. In 1949, a follow-up report authored by the radiologist Henry Garland from the University of California, San Francisco, demonstrated beyond doubt that 'reading the shadows' was a fickle affair.[14] Together, the reports presented proof of inter-individual and intra-individual variability – in about 30 and 20 percent of the cases respectively, readers differed from other readers or from their own previous evaluations. Researchers immediately realized the dangerous upshots: 'every day many persons throughout the country are being informed that their chests are free from disease when, in point of fact, they probably are not (and vice versa). This results in false security on the one hand and needless alarm on the other'.[15]

While there had been 'a tendency to assume that roentgenology is an exact science and that the objectivity of the medium defied error', the *Journal of the American Medical Association* editors commented that these 'astonishing' reports pointed to an incredible amount of inaccuracy made by trained X-ray observers.[16] These sensational findings on 'observer error' or 'observer variability', as researchers started calling these discrepancies, spurred a flurry of similar X-ray error investigations in a number of countries, including immediate replication studies in Denmark and the Netherlands.[17] Through the reports by Garland and others, it now appeared that radiologists had 'blind spots'. They had affinities for detecting particular types of lesions, needed (individually varying) eye-resting periods to be able to discern shadows, and could not agree whether a lesion should be classified as 'soft' or 'hard'. It even appeared that the 'attitude' of the observer (their 'optimistic' or 'pessimistic' outlook) might influence the interpretation. Many reports found similar percentages of variability, results that were of a 'disturbing magnitude', 'extremely disappointing', 'disheartening' and 'shocking' to researchers when they realized the ubiquity of what could now simply be called the 'error problem'.[18] It also appeared that there was no easy solution: when a group of highly experienced radiologists read and reread a set of survey films, they kept on arriving at the same degree of variability – a 'baffling' result.[19]

13 See Nicholas Binney, Christopher Hyde and Patrick M. Bossuyt, 'On the origin of sensitivity and specificity', *Annals of Internal Medicine* (2021) 174(3), pp. 401–7.

14 L. Henry Garland, 'On the scientific evaluation of diagnostic procedures', *Radiology* (1949) 52(3) pp. 309–28.

15 Garland, op. cit. (14), p. 325.

16 N.A., 'The "personal equation" in the interpretation of a chest roentgenogram', *Journal of the American Medical Association* (1947) 133(6), pp. 399–400.

17 For an overview of studies in different countries see N.A., *Abstracts of Papers Presented. Third International Congress of Photofluorography, Stockholm, Sweden, August 20–23, 1958*, Amsterdam: Excerpta Medica Foundation, 1958, p. 23. Researchers used concepts such as 'observer variation', 'accuracy', 'consistency', 'reliability', 'repeatability', 'reproducibility', 'identifiability' and 'detectability'. For an overview see Marcus J. Smith, *Error and Variation in Diagnostic Radiology*, Springfield: C.C. Thomas, 1967, 143–4, 157.

18 J. Yerushalmy, 'Reliability of chest radiography in the diagnosis of pulmonary lesions', *American Journal of Surgery* (1955) 89(1), pp. 231–40, 234.

19 L.H. Garland, 'On the reliability of roentgen survey procedures', *American Journal of Roentgenology and Radium Therapy* (1950) 64(1), pp. 32–41, 33.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

Healthcare professionals undertaking occupational radiological surveys were especially keen to assess reliability in reading thousands of images. A 1949 study in the *British Journal of Industrial Medicine* revealed serious incongruities examining survey images of coal miners' lungs in south Wales.[20] Readers could not agree which images should be regarded as normal, and which showed 'certifiable' pneumoconiosis (also known as black lung disease), which would merit compensation under the Workmen's Compensation Acts. Similarly, researchers assessing coal workers in France, Belgium, the Netherlands and the UK found many reader divergences and proposed a joint meeting to start an international system of standardized classification of chest X-rays.[21] In 1958, the International Congress on Medical Radiophotography included a separate section on observation errors, including presentations from Brazil, Finland, Romania and Poland on 'the human factor'.[22]

Through the 1950s, radiology became well known for its study of the observer error. This was not because errors of inconsistency and disagreement were unique to X-ray interpretation, but only because, as radiology researchers emphasized on various occasions, radiology lent itself more readily to quantitative evaluation of the degree of error and more precise data were available.[23] X-ray images provided suitably stable records – as one researcher remarked, not as 'flexible' as records of patient history, not as 'evanescent' as actual examination of the body – that could be subjected to multiple readings.[24] While the results were disconcerting, the project of quantifying error and proposing standards also afforded the discipline a certain objectivity. Researchers commended Garland for 'trying to lay a scientific foundation under roentgen diagnosis'.[25] Nevertheless, radiologists were often incredulous of the statistical evidence of their mistakes. Garland noted, 'One has to test oneself on a study of this kind to become fully aware of his own fallibility in this regard'.[26] As another researcher put it, 'only those who have themselves made duplicate readings of a series of films can come to appreciate the hard fact of their own unreliability'.[27]

This study of 'inherent error' in radiological observations spurred new scrutiny for the enduring problem of the 'personal equation' in scientific observation, or the 'human factor', as it was now more commonly called, pointing to a broader and long-standing epistemic problem of objectivity in medical science.[28] Inconsistencies in diagnostic observations were not limited to radiology but were made acutely visible through this particular disciplinary lens. Looking back at a decade of observer error research in 1959, Garland sketched a longer lineage of studies on error from the 1930s, which showed that medical professionals made mistakes in diagnosing emphysema, for example, or in the level of malnutrition in children. They erred in interpreting electrocardiograms and histologic

20 C. Fletcher and P.D. Oldham, 'Problem of consistent radiological diagnosis in coalminers' pneumoconiosis', *British Journal of Industrial Medicine* (1949) 6(3), pp. 168–83.

21 A.L. Cochrane, I. Davies and C.M. Fletcher, '"Entente radiologique" : a step towards international agreement on the classification of radiographs in pneumoconiosis', *British Journal of Industrial Medicine* (1951) 8(4), pp. 244–55. See Joseph Melling, 'Beyond a shadow of a doubt? Experts, lay knowledge, and the role of radiography in the diagnosis of silicosis in Britain, *c*.1919–1945', *Bulletin of the History of Medicine*, (2010) 84, pp. 424–66.

22 N.A., op. cit. (17), p. 23.

23 Paraphrasing Garland and Yerushalmy; Garland op. cit. (19), Yerushalmy, op. cit. (18).

24 Smith, op. cit. (17), p. 147.

25 Commentary by Robert R. Newell (father of Allan Newell) printed in Garland *et al*., op. cit. (19), p. 176.

26 Garland *et al*., op. cit. (19), p. 177.

27 Yerushalmy, op. cit. (18), p. 234.

28 'Inherent errors' in Garland, op. cit. (14), p. 324. On the longer genealogy of the 'personal equation' and observer errors see Jimena Canales, *A Tenth of a Second: A History*, Chicago: The University of Chicago Press, 2009. Rory Brinkmann, Andrew Turner and Scott H. Podolsky, 'The rise and fall of the "personal equation" in American and British medicine, 1855–1952', *Perspectives in Biology and Medicine* (2019) 62(1), pp. 41–71, 44.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

readings, in recording patients' medical histories and in counting red blood cells.[29] When more than fifty clinical laboratories were asked to test the same standard solutions they came back with different results.[30] A decade of recording errors by radiologists reframed these accounts from the past two decades as part of a common and pressing problem of observer variability, to which definite solutions had not yet emerged.

Variabilities in scientific observation had previously drawn the attention of philosophers and sociologists of science, notably Ludwik Fleck and Michael Polanyi. In 1935, Fleck, who was trained in microbiology, famously noted variabilities and uncertainties in scientific observation (puzzling views through the microscope, for example) to argue that individual observers were conditioned by a socially mediated thought style.[31] What could be observed depended upon observers' membership of collectives of researchers – 'thought collectives' – cultural constellations in particular. Polanyi in turn emphasized the role of tacit knowledge as an integral part of scientific knowledge formation, a learned understanding based on intuitive apprehensions that could not easily be articulated or formalized. The reading of chest X-rays aptly illustrated this latent dimension for Polanyi, who gained experience in evaluating such images as a medical officer during the First World War. Trained observers, like himself, could not but 'make sense' of such pictures, he argued; reading had become a form of 'personal knowledge'.[32] For Fleck and Polanyi, variability in observations could be understood by attending to social and individual learning processes that influenced processes of perception integral to scientific knowledge.[33] Yet the framework of 'observer error' emphasized in the 1950s foregrounded a subtly different epistemic attitude to scientific observation, focused on increasing accuracy by protocolizing, standardizing and formalizing X-ray reading. Radiologists framed observer variability as a problem – to which researchers proposed new technological and human solutions in the 1950s.

Reducing and taming errors: the imperfect radiologist as intuitive statistician

How were medical professionals to mitigate these inevitable errors in observation? One approach to the 'error problem' in medical practice that several mentioned was to start cultivating a greater attentiveness to the issue, for example by teaching courses on the 'factors affecting our judgement' to radiology students.[34] Yet from the mid-1950s onwards, other solutions started to come to the fore. The problem of observer error changed shape with and through developments in statistical theory, operations research, cognitive psychology and computer research, across a number of research sites and communities. Two Americans played a key role: Lee Lusted, a radiologist and radar specialist, and Robert Ledley, an engineer (specializing in dental prosthetics) and computer expert. Their work transformed the fallible trained observer, especially the radiologist, into a suboptimal medical decision maker who could be assisted by technology.

29 Studies listed in L. Henry Garland, 'The problem of observer error', *Bulletin of the New York Academy of Medicine* (1960) 36(9), pp. 570–84, 574.

30 Garland, op. cit. (29), p. 574, referring to a 1947 study by Belk and Sunderman.

31 Ludwik Fleck, *Genesis and Development of a Scientific Fact*, Chicago: the University of Chicago Press, 1981 (first published 1935).

32 Michael Polanyi, *Personal Knowledge: Towards a Post-critical Philosophy*, London: Routledge & Kegan Paul, 1958, p. 106.

33 Michael Hagner, 'Sehen, Gestalt und Erkenntnis im Zeitalter der Extreme: Zur historischen Epistemologie von Ludwik Fleck und Michael Polanyi', in Lena Bader, Martin Gaier and Falk Wolf (eds.), *Vergleichendes Sehen*, Munich: Wilhem Fink Verlag, 2010, pp. 575–95.

34 M.L. Johnson, 'A course on factors influencing scientific judgment', *Academic Medicine* (1955) 30(7), pp. 391–7; Johnson, 'Observer error: its bearing on teaching', *The Lancet* (1955) 2(6887), pp. 422–4.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

In the early 1950s, Lusted witnessed the 'error problem' at first hand. As a radiologist in training at the University of California in San Francisco, he was one of many volunteer X-ray readers in one of Garland and Yerushalmy's early 1950s observer variability investigations.[35] As he would later recount, these studies prompted him to investigate the possibility of technological solutions to improve interpretive accuracy in medical diagnosis. Lusted sought to implement in medicine his interests in computing and wartime expertise with engineering (radar) communication technologies. Historian Joseph November's incisive historical analysis of biomedical computing recounts how Ledley and the engineer Lusted started collaborating on a project of computerizing diagnosis, influenced by a shared background and interest in operations research, an applied science that brought a 'procedural rationality' to a range of disciplines from military missions to production processes and the development of early computer programs.[36] Both researchers believed that medical practitioners would greatly benefit from assistance by computers, for example in automatic data processing or calculating diagnostic probabilities. Yet in order to start building computers that could assist doctors, first the activities of doctors needed to be formalized; that is, redescribed in a potentially computable language. Ledley and Lusted modelled doctors' actions with future automation in mind.

The first and arguably most ambitious project they undertook was a formalization of the diagnostic reasoning process, published in the 1959 *Science* article 'Reasoning foundations of medical diagnosis', which would become widely cited and discussed.[37] This study provided mathematical descriptions of the complex reasoning process of diagnosis, a process at the basis of a doctor's 'feeling about the case'.[38] Should a five-week-old infant with throat tumours receive X-ray therapy or surgery? The authors divided this problem into separate parts, describing it in the language of logical equations, probability computations and statistics, as well as calculations they drew from decision analysis (referencing work on game theory and decision making by authors such as John von Neumann, Oskar Morgenstern, Duncan Luce and Howard Raiffa).[39] Applying this novel combination of mathematical techniques allowed for a separation of the 'strategy problem' (arriving at a medical diagnosis and the optimum strategy for treatment on the basis of probability calculations and statistics) and the 'values judgement problem' (calculating the trade-offs given certain moral, ethical, social and economic considerations). These techniques were meant to aid the physician, though the authors conceded that they also added new learning responsibilities. Physicians' tasks would become more complicated; they would have to study more and would need to be assisted by computers. Yet computers could never take over physicians' duties, the authors assured, they would simply make diagnosis more rigorous, i.e. more scientific.

Lusted and Ledley's 'logical analysis of medical diagnosis' redescribed the fallible observer as a medical professional involved in a complex diagnostic reasoning process. The minds of doctors were thought to work somewhat analogous to this proposed logical model: doctors seemed to perform computational tasks 'subconsciously' or on an

35 Lee B. Lusted, 'ROC recollected', *Medical Decision Making* (1984) 4(2), pp. 131–5.

36 Herbert A. Simon, *The Sciences of the Artificial*, Cambridge, MA: The MIT Press, 1969, pp. 9–23, 27. Cited in Joseph A. November, 'Early biomedical computing and the roots of evidence-based medicine', *IEEE Annals of the History of Computing* (2011) 33(2), pp. 9–23, 12.

37 Robert S. Ledley and Lee B. Lusted, 'Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason', *Science* (1959) 130(33), pp. 9–21.

38 Ledley and Lusted, op. cit. (37), p. 9.

39 John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press, 1944; R. Duncan Luce and Howard Raiffa, *Games and Decisions*, New York: John Wiley & Sons, 1957.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

'intuitive' level.[40] However, this intuitive diagnostic reasoning was also thought to be sub-optimal; doctors were underachieving in their logical capacities and in need of computational assistance.[41] In turn, the same mathematical techniques (calculations for probability, statistics and utility) thought to be at the foundation of the physicians' mind were also embodied in the (electronic) computational tools proposed to help them. Ledley and Lusted mentioned various (proto)types of diagnostic slide rule, mechanical correlator and punch card system, invented in France, the UK and the US since the mid-1950s, 'to assist the logical faculties' of doctors.[42] The historian Gerd Gigerenzer has described this analogical zigzag movement between *tools* (inferential statistical techniques as well as computational devices) and *theories* (of the workings of the (medical) mind) as a 'tools-to-theories' movement, characteristic of the period since the 1940s.[43] Statistical tools for testing hypotheses in a variety of inferential statistics approaches were considered 'in a new light as theories of cognitive processes in themselves'.[44] Looking at the doctor described by Ledley and Lusted in 1959 shows the important influence of this vision of the mind as 'intuitive statistician' in the 1950s.[45]

Radiologists, too, were thought to be intuitively reasoning experts. Developing logical principles for medical practice, Lusted proposed that radiology could serve as a 'testing ground' to further theorize and formalize various steps of this decision-making process and ultimately decrease observer error.[46] To do so, Lusted redescribed the actions of radiologists as a step-wise process: producing information on X-ray film, seeing the film (a physiological process), perceiving relevant aspects of the film, and diagnostic decision making. Each step could be formalized and also improved, at least hypothetically, by automation procedures for which Lusted mentioned some early prototypes. Radiologists' systematic decision making, for example, could be linked to the digital coding of X-rays. If the pattern of a tumour could be noted through a binary '1' and '0' code, visualized by black and white squares, this tumour profile could be read by someone who did not have any medical training, but merely needed to 'understand the code'.[47] In this pattern-reading example (based on an early prototype by radiologist Gwilym Lodwick in 1954), a 'coded analysis' was thought to enable an interpretation of a roentgenogram with the fewest errors, regardless of an observer's expertise.

Lusted was enthusiastic about these technological approaches to aid observation, even if evidence of increased accuracy was not yet in. Discussing a discontinued 1956 investigation into a screening device for mass miniature films – a pattern recognition apparatus to scan chest images and separate the normal from the abnormal – he noted that it was not the resulting machine that interested him but rather how the process of making it would require an understanding of the logic and probability principles underlying chest film interpretation. While an interest in new devices and procedures had brought Lusted to biomedical computing, ultimately the prospect of creating standards and

40 Ledley and Lusted, op. cit. (37), 10.

41 Lee B. Lusted, 'Logical analysis in roentgen diagnosis', *Radiology* (1960) 74(2), pp. 178–93, 178.

42 A number of mid-1950s examples, including F.A. Nash's diagnostic ruler and Ledley's own 1956 'logical aid', are listed in Ledley and Lusted, op. cit. (37).

43 Gerd Gigerenzer and David J. Murray, *Cognition as Intuitive Statistics*, London and New York: Psychology Press, 1987, p. 3.

44 Gigerenzer and Murray, op. cit. (43), p. xiii.

45 Gigerenzer and Murray, op. cit. (43), p. xiii. Marc Berg similarly draws on the tools-to-theory concept to understand the emergence of decision-support techniques in medicine, chiefly focusing on the period of the 1970s and 1980s. Marc Berg, *Rationalizing Medical Work: Decision-Support Techniques and Medical Practices*, Cambridge, MA: MIT Press, 1997.

46 Lee B. Lusted, 'Logical analysis in roentgen diagnosis', *Radiology* (1960) 74(2), pp. 178–93.

47 Lusted, op. cit. (46), p. 185.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

thereby also getting a 'firm grasp of the principles' behind observing patterns on X-ray images was most important.[48]

In Lusted's approach, which I have described as being at the forefront of the development of medical decision making as a field, we can discern two main and intersecting approaches to improving observer error. One was the aim to 'reduce error'. Even recognizing that not all mistakes could be eradicated, experimenters worked on designing more precise X-ray technologies and improving aspects of faulty diagnostic reasoning. Second, radiology researchers also approached diagnosis by what could be described as 'taming error', a strategy that regarded radiologists' false negative and false positive findings as unavoidable and aimed to monitor their relative occurrence.[49] The scanning device described by Lusted allowed researchers to calculate and determine an optimal 'operating point' on a statistical curve between too many false positives (warnings for lungs that were in fact normal) and too many false negatives (missed abnormalities).[50] Ultimately, a focus on 'error', both in reducing and in taming error, shaped a considerably positivist view of X-ray reading: the search for an optimal procedure to extract 'truth' from an image.[51]

These two intersecting approaches – reducing and taming error – also corresponded with two interconnected approaches to the radiology observer. First, the radiologist as 'intuitive statistician' seemed to be based on an individualizing approach, viewed as a single mind calculating probabilities and trade-offs. Yet on second view, individual doctor's observations went beyond the singular, since in Ledley and Lusted's 1959 vision each probability calculation of an individual case would feed into a data collection of the most current statistics. Beyond the diagnostic punch card aids of the mid-1950s, Ledley and Lusted now imagined something much bigger: a 'central health computing and records service'. They envisioned a data-sharing network between local hospital computers and a central research computer, through which the central node would continuously be fed with new statistics and automatically drop older ones, allowing for calculations on the basis of the most current trends – the computer would 'learn by experience'.[52] With this vision of a networked diagnostic calculation model, Ledley and Lusted had connected individual patient diagnosis to a population scale.

Although widely noted, Ledley and Lusted's logical and technological model had hardly solved the ever-present problem of variability in interpreting X-ray images. Around 1960, new studies showed that observer variability remained a problem, especially in mass survey work.[53] Beyond 'logical analysis', researchers involved in mass X-ray imaging were looking for concrete measures to improve the accuracy of their procedures reading

48 Lusted, op. cit. (46), p. 185. November, op. cit. (36), p. 94, notes that Lusted's vision for improving medicine was not the use of computers per se, but a change towards a shared commitment to standards and quantifying techniques that would ultimately result in a common quantitative language for complex medical information.

49 My use of 'taming error' is inspired by the work of Claudia Aradau and Tobias Blanke, who draw on Ian Hacking's 'taming chance'. Claudia Aradau and Tobias Blanke, 'Algorithmic surveillance and the political life of error', *Journal for the History of Knowledge* (2020) 2(1), pp. 1–10.

50 Towards the end of the 1960s, radiology researchers started employing 'receiver operating curves' (ROC curves, derived from signal detection theory) as a way to measure (and conceive of) optimal diagnostic performance. See Gigerenzer and Murray, op. cit. (43), pp. 42–60.

51 Conceptually, a focus on error drew attention away from the impossibility of a diagnostic 'ground truth' in radiology. At times, radiology researchers did acknowledge an inevitable element of uncertainty in reading images. For example, in his introduction to *Error and Variation in Diagnostic Radiology* Smith defines error as 'a wandering from the truth (even though the truth may not be known)'. Smith, op. cit. (17), p. 5.

52 Ledley and Lusted, op. cit. (37).

53 For example, Smith, op. cit. (17); C. Wegelius, 'Röntgenreihenuntersuchungen mit dem Schirmbildverfahren', in H. Vieten (ed.), *Allgemeine Röntgendiagnostische Methodik*, Berlin: Springer Verlag, 1966, pp. 600–39.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

hundreds of thousands of images. At the scientific department of the UK National Coal Board, for example, researchers aimed to derive a 'quantitative measure of the accuracy of the reading process'.[54] Attempting a precise mathematical description of the process of recognition and interpretation of the X-ray images would be 'impracticable', they explained.[55] Instead, they were looking for tangible directions: when could a common reading be taken as 'definitive'? Would a second or even a third reading increase diagnostic accuracy? Was it possible to devise a practical, human solution to the 'human factor'?

Doubling the fallible trained observer: dual X-ray reading and negotiating expertise

The problem of observation variability proved tenacious. Even if expert observers had plenty of time for perceiving and interpreting, and were well-rested and provided with the most precise images, a considerable number of disagreements and inconsistencies persisted. Garland's first report had already suggested a simple potential solution. Following up in 1950, Yerushalmy examined 'dual reading' as a way to decrease error.[56] Also called 'double reading', this meant performing a second independent interpretation of an X-ray completely separate in time from the first reading, either by a second observer or by the same observer on a second occasion. While there was a danger that second opinions would merely multiply the errors (more false positives and false negatives), Yerushalmy's research demonstrated that dual reading decreased errors. It was also cost-efficient, he reasoned, because the expenses of a missed case would be much greater than the costs of multiple readings.[57] Subsequent studies in various countries predominantly agreed with these findings.[58] However, the procedure did not seem to be implemented widely. Writing in *The Lancet* in 1955, two UK radiologists lamented, 'Nearly nine years have elapsed since the presentation of the first paper on this problem, and the chief conclusion – the importance of a second reading – is still neglected in this country'.[59]

The reasons for the puzzling failure to take up what seemed a sensible human solution to the fallible observer are multifaceted and I want to speculate on a number of them. First, the prospect of increasing – possibly almost doubling – the observer workload may have discouraged many professionals for logistical and economic reasons.[60] Moreover, increasing the number of workers also spotlighted the thorny issue of expertise in the X-raying workforce. Because even expert radiologists now appeared to be fallible observers, the 'error problem' drew heightened attention to the evaluation of individual performance and also to the demarcation of radiological expertise. Reports on observer

54 J.W.J. Fay and J.R. Ashford, 'The study of observer variation in the radiological classification of pneumoconiosis', *Occupational and Environmental Medicine* (1960) 17(4), pp. 279–92, 280.

55 Fay and Ashford, op. cit. (54), p. 280.

56 J. Yerushalmy *et al*., 'The role of dual reading in mass radiography', *American Review of Tuberculosis* (1950) 61 (4), pp. 443–64.

57 J. Yerushalmy, 'Problems in radiological interpretation', *California Medicine* (1949) 70(1), pp. 26–30, 29.

58 Some examples of early studies that pointed to the efficacy of dual reading are E. Groth-Petersen, A. Lovgreen and J. Thillemann, 'On the reliability of the reading of photo-fluorograms and the value of dual reading', *Acta Tuberculosea Scandinavica* (1952) 26(1–2), pp. 13–37; W.A. Griep, 'The role of experience in the reading of photofluorograms', *Tubercle* (1955) 36(9), pp. 283–6.

59 P. Stradling and R.N. Johnston, 'Reducing observer error in a 70-Mm. chest radiography service for general practitioners', *Lancet* (1955), 268(6877), pp. 1247–50, 1249.

60 In 1969, William Tuddenham simply remarked, 'the procedure has never appeared economically feasible in routine practice of radiology'. W.J. Tuddenham, 'Roentgen image perception: a personal survey of the problem', *Radiologic Clinics of North America* (1969) 7(3), pp. 499–501, 500. In the 1950s, routine double reading seems to have remained an exception; see Danish Tuberculosis Index, 'Dual reading as a routine procedure in mass radiography', *Bulletin of the World Health Organization* (1955) 12(1–2), pp. 247–59.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

error frequently showed a score of unreliability between individual trained radiologists or groups of experts (for example, a study by Garland and A.L. Cochrane compared American and British experts) as well as between experts and other, non-radiology trained readers, such as chest experts or swiftly trained-up mass-survey readers.[61]

While some researchers regarded lower-skilled readers as too unreliable, others emphasized that with adequate training, the proper level of experience could be obtained quite easily. Dutch tuberculosis researcher W.R. Griep argued in 1955 that inexperienced chest experts could reach a 'maximum reliability' after about three years of training.[62] Perhaps such readers-in-training could even start to work almost immediately: his research showed that a procedure of dual reading with two inexperienced readers could bring observer errors to a level equal to two experienced readers. Maybe, Griep speculated, it was not experience that mattered most, but the reader's character. Drawing conclusions from a test of (merely) five observers, he wondered whether the numbers of over-reading (erroneously calling an image suspicious, i.e. false positive) might have to do with the character of the female specialists tested for this investigation. While the male reader 'does dare to decide if an affection he sees is important or not', he noted, the two women exhibited an attitude that 'you never can tell'.[63] Though Griep's invocation of a gendered aspect in male courage versus female caution is just a fleeting remark, it is telling for a broader negotiation of (X-ray) reading expertise in light of non-physician and women workers entering new (or shifting divisions of) tasks in medical and laboratory fields.

While women were already ubiquitously employed in mass X-ray survey work and radiology clinics as clerks and radiographers (operators who worked with patients and machines to produce the X-ray images), in the mid-twentieth century non-physician readers, including women, were increasingly considered a potential source of cheap labour to serve expanding survey facilities.[64] Outside radiology, other disciplines similarly grappled with a demand for workers who could process and read massive numbers of images. For example, in novel population-wide campaigns to detect cervical cancer, a new division of mostly female 'screeners' was educated to catch suspicious pap smears among thousands of microscopic slides.[65] In physics research, 'scanning girls' were trained to handle and interpret large numbers of particle track images.[66] In both examples, reading work was modelled (and discursively framed, i.e. 'feminized') to align with an affordable female workforce by restructuring work processes and continuous performance evaluation.[67] In contrast, in the field of (survey) radiology, the rise of non-physician readers remained contentious. Radiologists continued to protect their professional ownership of diagnosis against replacement by paramedical reading personnel. While the need for more and possibly cheaper non-physician personnel was frequently voiced – especially in the context of

61 A.L. Cochrane and L.H. Garland, 'Observer error in the interpretation of chest films: an international investigation', *The Lancet* (1952) 260(6733), pp. 505–9; Garland, op. cit. (19), p. 35.

62 Griep, op. cit. (58).

63 Griep, op. cit. (58), p. 285.

64 For example, in 1952, the American Society of X-Ray Technicians had four thousand members, about 75 per cent of whom were women. Mildred Barber, *The Outlook for Women as Medical X-ray Technicians*, Bulletin of the Women's Bureau No. 203–8, Medical Services Series, 1954, p. 36.

65 Monica Casper and Adele Clarke, 'Making the pap smear into the "right tool" for the job', *Social Studies of Science* (1998) 28(2), pp. 255–90.

66 Peter Galison, *Image and Logic: A Material Culture of Microphysics*, Chicago: The University of Chicago Press, 1997.

67 The shaping of image processing and reading as feminized work meant that this labour was presented (to a varying extent) as tedious and deskilled and eventually replaceable by machines. See Casper and Clarke, op. cit. (65); and Galison, op. cit. (66). On feminized work in the mid-twentieth century see Mar Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*, Cambridge, MA: MIT Press, 2018.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

a call for double reading – the practice of interpreting X-rays was demarcated as the task of a medical specialist.[68]

Economic and sociocultural considerations about the cost-efficiency of medicine and the esteem of non-physician (including women) workers thus influenced how solutions to observer variability were conceived and realized, and dual reading could not be aligned with principles and practices in radiology and medicine in the 1950s. Yet I want to suggest that there was another, arguably more fundamental, reason why the practice was not taken up. Dual reading emerged as a solution from an experimental and statistical framing of radiological practice, shaped by a 'laboratory' imitation of what scientific (medical) observers do. In the 1950s, the experiments by Garland, Yerushalmy and other investigators simulated X-ray viewing practices under experimental conditions, and emphasized recording individual researchers' interpretations, providing numbers that would feed into statistics of variability. This experimental frame could not fully describe, however, the way mass X-ray workers viewed and interpreted images in messy real-world situations.

Dual-reading research makes this discrepancy between model and real-world, laboratory investigations and actual readers looking at pictures of lungs sharply visible. In 1960, researchers working at the National Coal Board aimed to outline uniform procedures for taking and reading radiographs.[69] With mass-X ray examinations of workers at no less than twenty-five coalmines, standardized methods were necessary to produce accurate data to investigate the progression of black lung disease.[70] Attempting to keep observer variability in check, the researchers proposed a triple reading process (dual reading by one medical officer, and a third reading by another). While the statistical analysis of 'dual reading' required that a first and second reading be wholly independent, the model could not contain the fact that doctors tended to remember a previously seen image – especially the 'doubtful' films – even several months after the fact. 'Double reading' hardly matched actual reading practice. Reflecting on their experimental model, the researchers commented, 'it is apparent that it does not provide an entirely realistic representation of the reading process on any particular film'.[71] To reduce observer error, researchers started investigating models of 'joint discussion', which appeared to reduce inconsistencies better than separate dual readings between different observers.[72] Gradually, variants of collegial joint discussion about uncertain images were strengthened and reframed as forms of 'conference reading' taking place in 'referee conferences' and became more explicitly implemented in working routines.[73] Such practices did not amount to the fully fledged programme of dual reading proposed as a solution to the problem of 'observer error'. Instead, practices mitigating uncertainty had already evolved outside the investigative experimental sites of radiology research from which the paradigm of the 'error problem' emerged.

68 Towards the end of the 1960s, new systematic investigations into (perceived) lower-skilled (often women) readers of X-rays merged in tandem with novel mass X-ray survey campaigns for breast cancer research. See Franklin S. Alcorn and Evelyn O'Donnell, 'Mammogram screeners: modified program learning for nonradiologic personnel', *Radiology* (1968) 90(2), pp. 336–8.

69 Fay and Ashford, op. cit. (54).

70 Fay and Ashford, op. cit. (54); See J.W.J. Fay, 'The National Coal Board's pneumoconiosis field research', *Nature* (1957) 180(4581), pp. 309–11.

71 Fay and Ashford, op. cit. (54), p. 283.

72 See references in Fay and Ashford, op. cit. (54).

73 'Conference reading' is mentioned as a common practice in Esmond Ray Long and Seymour Jablon, *Tuberculosis in the Army of the United States in World War II: An Epidemiological Study with an Evaluation of X-Ray Screening*, Washington DC: US Government Printing Office, 1955. 'Referee conferences' are recommended in 'Report of the Joint Committee on chest-X-ray', *California Medicine* (1954) 80(4), pp. 343–4.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

The recurring discovery of the error problem

Through the 'error problem', as it was shaped in the field of radiology in the mid-twentieth century, a new epistemic position was foregrounded: the expert as a fallible trained observer. This expert needed to be monitored and assisted to reduce mistakes and stay within a statistical range of acceptable inconsistencies and disagreements, prompting a search for different measures and aids to discipline error. My analysis of these faulty X-ray readers contributes to historical research on scientific observers and historical epistemology in the mid-twentieth century. Historians Lorraine Daston and Peter Galison have pointed to the emergence of a new epistemological position in the 1930s: a focus on 'trained judgement' in the interpretation of visual records in scientific practice by the 'trained expert' who has developed a capacity 'to synthesize, highlight, and grasp relationships in ways that were not reducible to mechanical procedure'.[74] In Daston and Galison's account, the early decades of the century are characterized by a shift from a focus on creating scientific graphs and images conceived as having a self-evidential nature according to an ideal of 'mechanical objectivity' towards an emphasis on the necessity of the trained eyes of experts to identify and judge the characteristics of these records. Daston and Galison describe the 'trained expert' as someone who 'embraced instruments, along with shareable data and images, as the infrastructure on which judgment would rest'.[75] My genealogy of the imperfect radiologist and the fallible trained observer shows that some trained experts were viewed as inevitably in need of help, and notions of judgement were themselves increasingly shaped in terms of statistical analyses that counted variations between observers, beyond the individual trained expert.[76]

At the turn of the 1940s, the notion of the imperfect trained observer took shape with and through the statistical monitoring of observer errors in experimental (investigative) set-ups. I have argued that this framework also shaped how solutions to the error problem, such as technological aids for automated image recognition and procedural changes such as 'dual reading', could be imagined. Comments on inter-observer variability were not new in the 1940s, but refracted earlier observations on scientific observations by philosophers and sociologists of science Ludwik Fleck and Michael Polanyi, who considered skilful, intersubjective, intuitive and unaccountable elements as integral dimensions of scientific observation. In contrast, I have shown that the framework of the fallible observer prompted a statistical and experimental solution: a second reader could potentially decrease error. Yet research in dual reading hardly modelled the reality of mass X-ray observations and radiological practice, obscuring a more complex reality in which readers could jointly discuss what they saw on the image and, as Fleck and Polanyi suggested, draw on a complex process of collective training and experience. My analysis also suggests that despite the repeated lament that dual reading had not been implemented, there were already collaborative practices in radiology, hiding in the shadows.

Today, the figure of the fallible expert observer has regained significance. Advocates of AI in radiology are refocusing attention on the persistent issue of reader variability and propose artificial intelligence as a promising and fitting technological solution to this issue of human error.[77] On second view, however, this 'rediscovery' of radiology's error

74 Lorraine Daston and Peter Galison, *Objectivity*, Cambridge, MA: MIT Press, 2007, p. 314.

75 Daston and Galison, op. cit. (74), p. 329.

76 My observations on a conceptual move from individual judgement towards the importance of statistical guidelines are much in line with the concept of 'regulatory objectivity' in Alberto Cambrosio, Peter Keating, Thomas Schlich and George Weisz, 'Regulatory objectivity and the generation and management of evidence in medicine', *Social Science & Medicine* (2006), 63(1), pp. 189–99, 190.

77 For example, Elizabeth Krupinski, 'Artificial intelligence: lessons learned from radiology', *Healthcare Transformation*, December 2019, pp. 5–10.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

problem is not new but is another instance of its continuous return. Each decade, so it seems, a group of researchers revisits the 'Achilles heel' of the discipline, to point back to the first reports around 1950 and conclude that the problem of human error in radiological observation is multifaceted, complex and persistent.[78] Over the past seventy-five years or so, the recurring recognition of the error problem is closely tied to returning suggestions for reducing error.

My historical analysis shows how, in the decade after the first reports on inter-observer variability, technological solutions to automate X-ray reading may have received as much attention as procedural solutions to mitigate error, if not more, which implies increasing and restructuring human X-ray reading labour. After the 1950s, novel computer-assisted procedures for X-ray reading (based on processes of standardization and automated pattern recognition) have continued to be much publicized, from computer-aided analysis of X-ray images in the 1960s to computer-aided diagnosis (CAD), computer-aided detection (CADx) and the use of artificial neural networks from the 1970s to the 2000s. By looking back to the early days of the debate on improving X-ray image reading, my analysis helps us to understand how the promise of a technological solution to a pressing error problem could be sustained vis-à-vis (an allegedly) daunting and overly expensive human fix. Considerably less publicized is the associated recurring discovery of dual reading. A recent (2018) meta-review again suggests that double reading in radiology is beneficial, but also adds a familiar caveat: 'the benefit of double reading must be balanced by the considerable number of working hours a systematic double reading scheme requires'.[79]

Historicizing the paradigm of the 'error problem' reveals the rhetorical negotiation between two poles – an up-and-coming image-reading technology in relation to a suboptimal human. My analysis also opens up a view of the way this pairing has taken a new turn: in the 1990s, computer-aided detection started to be proposed as a way to realize double-reading procedures, particularly in the context of mass screening. At that point, CAD became envisioned as 'viable cost-effective alternative to double reading by radiologists'.[80] With strategic modesty, technology was positioned as the not-yet-perfected but feasible assistant to the inevitably imperfect X-ray-reading expert. 'AI as second reader' is the present-day version of this diplomatic conceptual emplacement, which has helped to sustain interest in computer-aided and AI-supported solutions at a time when evidence that such methods improve clinical accuracy and efficacy is still pending.[81] This vision of AI as a potentially cheaper double reader, I demonstrate, has long been in the making.

My historical account of observer errors in the field of X-ray reading helps to contextualize the bombastic contemporary trope of the 'centaur radiologist', mentioned at the beginning of this article. Half-man and half-AI, the centaur radiologist conveys the image of an effective doctor–warrior that harmoniously combines human skilfulness with the newest AI technologies to 'find patterns in data that are beyond humans' abilities'.[82] As I show, this heroic image obscures a more mundane and modest application

78 P.J. Robinson, 'Radiology's Achilles' heel: error and variation in the interpretation of the röntgen image', *British Journal of Radiology* (1997) 70(839), pp. 1085–98.

79 Håkan Geijer and Mats, Geijer 'Added value of double reading in diagnostic radiology,a systematic review', *Insights into Imaging* (2018) 9, pp. 287–296, 296.

80 Heang-Ping Chang, Shih-Chung B. Lo, Berkman Sahiner, Kwok Leung Lam and Mark A. Helvie, 'Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network', *Medical Physics* (1995) 22(10), pp. 1555–67, 1555.

81 Current AI-supported reading developments in radiology may justify expectations of future clinical efficacy, some argue. See, for example, Luke Oakden-Rayner, 'The rebirth of CAD: how is modern AI different from the CAD we know?', *Radiology: Artificial Intelligence* (2019) 1(3), e180089.

82 Keith J. Dreyer and J. Raymond Geis, 'When machines think: radiology's next frontier', *Radiology* (2017) 285 (3), pp. 713–18, 714.

https://doi.org/10.1017/bjt.2023.5 Published online by Cambridge University Press

of AI as an allegedly more affordable aid in optimizing the reduction of errors. At the centre of this development, starting in the late 1940s, is the imagination of a fallible radiological expert.

Acknowledgements. I would like to thank the editor and two anonymous referees of the *BJHS* for their insightful reviews of earlier drafts. Special thanks to Richard Staley and the editors of this issue on Histories of AI for their expert conceptual and editorial advice and for welcoming me at the Mellon Sawyer Seminar on Histories of AI at the University of Cambridge in 2021. This article has also benefited from generous reflections on a draft by members of the Maastricht University Science, Technology and Society Studies research group – special thanks to Joeri Bruyninckx for his careful comments. Additionally, I would like to thank the radiologists at the Trefpunt Medische Geschiedenis Nederland, in particular Kees Simon and Frans Zonneveld, who responded kindly, swiftly and extensively to my questions and provided vital professional reflections. I thank Eddy Houwaart for important feedback in early stages of developing this text. Thank you Alex Campolo for perceptive exchange on notions of optimization and error. My research was generously supported by the Dutch Research Council (NWO) as part of the RAIDIO research project grant (number 406.DI.19.089). Special thanks to RAIDIO team member Sally Wyatt for important advice on drafts, as well as to Annelien Bredenoord, Jojanneke Drogt, Karin Jongsma, Megan Milota and Shoko Vos.

Cite this article: Lysen F (2023). Errors and fallibility in radiology: X-ray readings and expert radiologists, 1947–1960. *BJHS Themes* **8**, 127–141. https://doi.org/10.1017/bjt.2023.5