The Challenges of Identifying Significant Epistemic Failure in Science

Tobias Lehmann, Michael Borggräfe, and Jochen Gläser

1. Epistemologically Cherished but Pragmatically Ignored

If one follows the accounts by philosophers of science and the discussions in scientific communities, there can be little doubt that failure is an essential part of scientific practice. It is essential both in the sense of being integral to scientific practice and of being necessary for its overall success. Researchers who create new scientific knowledge face uncertainties about the nature of the problem they are trying to solve, the existence of a solution to that problem, the way in which a solution can be found, and their ability to find such a solution (Gläser, 2007, 247). The existence of these uncertainties means, in turn, that each research process carries the risk of failure.

The inseparability of success and failure is the reason why the latter is important from an epistemological perspective. In the philosophy of science, Popper (1992 [1935]) argued that scientific progress depends not on the verification of proposed hypotheses, but rather on the opposite mechanism of their falsification. In this sense, any hypothesis, or broader theory, that is to be called scientific must provide the possibility of its own failure. This argument resonates in scientific communities (Mulkay/Gilbert, 1981; Sovacool, 2005), whose members emphasise the importance of failure (Clark/Thompson, 2013), criticise the neglect of negative findings and argue that the publication of negative results and results contradicting other researchers' findings merits special attention by editors and authors (e.g. Pfeffer/Olsen, 2002; Loscalzo, 2014; Fraser et al., 2018).

This epistemological admiration is curiously at odds with the practice of scientific communities. In their everyday knowledge production and publication activities, scientific communities tend to ignore and even suppress reports on failure. A "publication bias" – a tendency to publish only positive results – has been observed in several disciplines (Sterling, 1959; Stanley, 2005; Dwan et al., 2008; Fanelli, 2012; Franco et al., 2014). This bias appears to exist because all parties involved in the publication process – authors, journal editors and reviewers – tend to prefer the publication of positive results. If negative results are published, they take longer time to publication in comparison to studies reporting positive results (Stern/Simes, 1997; Misakian/Bero, 1998; Suñé et al., 2013) and receive fewer citations than positive results (Gumpenberger et al.,

2013). The latter effect is field-specific in that this citation penalty seems most prevalent in the biological and biomedical sciences and least prevalent in the physical sciences (Fanelli, 2013; Duyx et al., 2017).

Given this odd triangle of empirical ubiquity, epistemological admiration and practical neglect, it is surprising that scientific failure has been largely ignored by science studies. Neither historical reconstructions of prominent cases of failure nor observations of failure in ethnographies have led to systematic theoretical accounts that link forms of failure and the conditions under which they occur to specific effects. This is unfortunate because failure in science is an important site for the investigation of mechanisms of knowledge production in scientific communities. Failure disrupts individual and collective production processes and thus has the potential to expose taken-for-granted assumptions about how they work. Situations in which mechanisms of knowledge production do not produce the expected outcomes force researchers and their communities to cope with unexpected changes in their knowledge production. Studying these situations and responses is likely to contribute to our understanding of the ways in which scientific communities produce knowledge the same way the study of infrastructure breakdowns contributes to the understanding of infrastructure: "[t]he normally invisible quality of working infrastructure becomes visible when it breaks" (Star, 1999, 382).

Another reason why scientific failure deserves more attention is its increasing relevance to the continuation of academic careers. Failure limits opportunities to publish and thus to gain visibility and reputation. This is likely to induce risk avoidance in the selection of problems and approaches to solving them, which in turn may slow down the progress of knowledge production. Studying career implications of failure can therefore lead to a better understanding of the processes by which the production and communication of contributions are linked to the development of researchers' reputation and to the translation of this reputation into career advancement.

Paradoxically, studying scientific failure could help us understand scientific success. This makes it all the more troubling that failure is largely ignored. One reason for this neglect is the difficult empirical access to the phenomenon. Since failure often prevents publication, and since publications of failure are often ignored in the communication processes of scientific communities, it leaves few traces for outside observers.

Against this background, this chapter presents conceptual and methodological considerations of the project "Forms and Effects of Failure in Science" (FEM), which we currently conduct at the Technical University of Berlin.[1]

1 For the project website, see: https://fem.hypotheses.org.

The project focuses on field-specific forms of significant epistemic failure, i.e. when researchers do not achieve the goals of their research processes or research programmes. It addresses the consequences of these forms of failure for the careers of the involved researchers and for the knowledge production of their scientific communities. In this chapter, we discuss the state of research on scientific failure (2) and present our theoretical approach to the conceptualisation of significant epistemic failure as socially constructed by researchers and their scientific communities (3). We then discuss our approach to the identification of candidate cases for significant epistemic failure for an interview-based study of forms and effects of failure (4). Our attempts to empirically identify research processes that may have failed enable some conclusions on the construction of (field-specific) forms of significant epistemic failure (5).

2. Isolated Case Studies and Undertheorised Observations

Given its ubiquity in the production of scientific knowledge, it is not surprising that failure has been observed, and sometimes investigated, by science studies. However, it has rarely been addressed as such, and has not been the target of theory building. Laboratory ethnographers, for example, have observed instances of failure in everyday laboratory work and the ways in which researchers cope with them (Knorr-Cetina, 1981, 52–53, 63–66; Lynch, 1985, 81–140; Latour/Woolgar, 1986 [1979], 129–141, 154–155). However, they used their observations mainly as building blocks for arguments about the constructed nature of scientific knowledge. The attempt by Star and Gerson (1987) to combine some forms of everyday failure (mistakes and artefacts) with fraud and discovery in a generalised account of "anomalies" that need to be "managed" by scientists was not developed further.

Some cases of failure were considered instructive and have been investigated in detailed case studies by both the history and the sociology of science. These include the 'discovery' of N-rays and the dismissal of this discovery by the scientific community (Nye, 1980), the fate of Pflüger's theory of diabetes (Schlich, 1993), and the 'discovery' of cold fusion, which attracted researchers and sponsors for a long time after it was dismissed by the scientific community (Simon, 1999). None of these single-case studies offered a systematic account of their case as an instance of failure, i.e. as a member of a specific class of phenomena. The case study by Collins (1975; 1981; 1999) of Weber's claim to have detected gravitational waves, which was not accepted by the scientific community, informed theoretical considerations concerning controversies in science but did not lead to theoretical approaches to scientific failure.

These case studies illustrate that scientific communities can explicitly – by dismissing a researcher's claim – or implicitly – by ignoring it – construct the failure of a research process even though researchers themselves consider their work as successful. The resistance of researchers to new scientific findings, which was discussed by Barber (1961) and by Kuhn (1962) in his account of scientific revolutions, can also amount to a community's construction of research processes as failed.

Merton's work on priority disputes (1957; see also Cozzens, 1989) and Hagstrom's (1974) observation of scientists' experiences of being anticipated by colleagues point to a particular kind of failure, namely the failure to make a novel contribution. In these cases, the failure is not related to the content of a contribution but to the community's construction of the contribution as obsolete. The production of scientific knowledge is a 'winner-takes-all' competition in which only the first contribution of a particular knowledge claim is used by the scientific community (Dasgupta/David, 1994). If similar scientific findings are presented to the community at the same time, the community needs to decide whose contribution it uses, which in some cases leads to priority disputes between researchers because reputation is allocated accordingly (Merton, 1957; Cozzens, 1989). If researchers observe others publishing a result they were still trying to produce, they are anticipated by others (Hagstrom, 1974). These studies again highlight the role of scientific communities, which construct one researcher's contribution as relevant while disregarding others.

Several social phenomena that are relevant as partial causes of failure have been studied but have not been integrated in systematic accounts of failure. Historians studied the role of experimental error in research processes, e.g. in the case of Johannes Kepler (Hon, 1987) or Louis Pasteur (Cadeddu, 2000). Hon (1989) developed a typology of experimental error in the stages of an idealised research process on the basis of historical cases, and distinguished between mistakes and errors (Hon, 1995; for an application to science in general see Schuol, 2020). However, it remains unclear which kinds of experimental mistakes and errors may or may not produce failure, under what conditions this might happen, and to what effects.

Effects of failure have been studied both ethnographically and bibliometrically. Ethnographic studies have addressed researchers' personal experiences and responses to failure, but did not explore structural consequences for their careers (Timmermans, 2011; Sigl, 2016). In a similar vein, contributions to a volume edited by Fam and O'Rourke (2020) document personal perspectives on the failure of interdisciplinary and transdisciplinary research projects, which are of more empirical than conceptual value.

Two strands of bibliometric research address effects of failure. Citation studies of publications of negative results have identified a field-specific 'citation

penalty' for such results (Fanelli, 2013; Duyx et al., 2017). These findings suggest that negative results might offer fewer opportunities to build on them in subsequent knowledge production processes, a possibility which has not yet been explored. Bibliometric studies also identified cases in which publications were considered as not containing relevant contributions immediately after they appeared but were reassessed later. These so-called "sleeping beauties" (van Raan, 2004) appear to be rare but demonstrate the dynamics of the community's assessments of contributions. One cause of the under-appreciation of contributions seems to be that they are ahead of their time, i.e. that the research context in which they constitute a relevant contribution does not yet exist at the time of their publication. These situations are discussed as "premature discoveries" (Hook, 2002).

This account of the literature suggests that scientific failure is indeed a common occurrence in the laboratory (according to ethnographic studies) and that failed research processes may have a significant impact on a scientific community's knowledge production (according to case studies). However, ethnographic observations, (historical) case studies, discussions of potential causes of failure and bibliometric studies of effects of failure have not led to sociological middle-range theories, i.e. to accounts that explain how different types of failure with particular effects occur under specific circumstances. The occurrence of particular forms of failure appears to be difficult to predict, and failure is often invisible due to its neglect in the communication processes of scientific communities. In our FEM project we are taking up this challenge.

3. Theoretical Considerations: The Social Construction of Significant Epistemic Failure

In order to investigate forms and effects of scientific failure, we need to define it and to identify the particular forms of scientific failure we are interested in. We apply an action-theoretical perspective on failure and define it as a situation in which actors do not achieve the goals of their action. This definition is ideal-typical because actors may (and often do) respond to the perception or anticipation of failure by modifying their goals. Nevertheless, there are also situations in which the goals are clearly not achieved, and which cannot be 'fixed' by further adaptation. In the case of research these include, for instance, the above-mentioned situation in which gravitational waves were not detected by Weber and the situation in which cold fusion could not be experimentally demonstrated.

In the following, we discuss the application of this definition to scientific failure and introduce three specifications. First, we focus on *epistemic failure*,

i.e. failing practices of knowledge production. Second, we distinguish between two forms of epistemic failure, namely epistemic failure in everyday scientific work and *significant epistemic failure*, which we understand as researchers not achieving the goals of a research process or research programme (3.1). Third, we ask how, by whom and in which frame of reference the (non-)achievement of goals is constructed (3.2).

3.1 *Significant Epistemic Failure as Failure to Achieve Goals of Research Processes or Research Programmes*

We focus on significant epistemic failure, which we define as a situation in which researchers fail to achieve the goals of their research processes. Research processes are single or composite actions aimed at (a) producing knowledge claims that (b) close specified gaps in the community's knowledge. They can be distinguished from operations, which are parts of actions and whose purpose and meaning is derived from the actions they are part of.

The distinction between operations and actions supports the delineation of significant epistemic failure. Uncertainties inherent to scientific knowledge production as well as mistakes made in everyday work produce a wide spectrum of things that can go wrong. These range from the 'everyday messiness' of laboratory work in small-scale operations such as the calibration of instruments, the preparation of samples or the conduct of measurements to the production of disappointing results from single experiments and to the failure of particular experimental approaches. Laboratory studies have demonstrated that small failures like these occur constantly. They can often be 'fixed' or 'worked around' in the course of a research process, or they may initiate changes of plans, without endangering the achievement of goals. In contrast to this fluid development of everyday failures and adaptation, we consider epistemic failure as significant if research processes or even comprehensive research programmes (i.e. series of research processes with the same thematic focus) do not achieve their goals. The significant epistemic failure we are interested in is marked by the observation at the end of a research process that the research goals of a project (or programme) have not been achieved, and that no more steps can be undertaken to 'fix' the situation.

3.2 *Significant Epistemic Failure as a Multi-Level Phenomenon of Social Construction*

The success or failure of research processes is the outcome of a process of social construction. Significant failure can be constructed by any observer who is able to assess the congruence of goals and results of a research process: congruence means success, non-congruence means failure. The literature suggests

that success or failure of research processes are constructed by the researchers themselves and by their scientific communities.[2] Scientific knowledge is not produced by single researchers or research groups but by scientific communities whose members jointly develop a shared body of knowledge by interpreting it, constructing knowledge gaps out of this interpretation, and offering contributions that close these gaps in publications. If these offers are used by other members of the community, they are integrated into the shared body of knowledge (Gläser, 2006). From this follows that the achievement of goals by community members is evaluated by the members who offer new knowledge claims and by their scientific community, whose members use or neglect the offered contributions.

These evaluations occur in two frames of reference. A first frame contextualises the epistemic content of a knowledge claim. In this frame of reference, researchers and their communities consider whether a *specific gap in the community's knowledge has indeed been closed* by the research process. They establish whether a knowledge claim exists and assess its validity and reliability. In a second, more abstract, frame of reference researchers and their communities consider whether a *relevant contribution to the community's knowledge has been made*, i.e. whether a gap in the community's knowledge whose closure is relevant to the community's further production of knowledge has been identified and closed. The difference between both frames of reference is illustrated by the many contributions that are ignored by the scientific community even though they are considered to be outcomes of successful research processes by authors, journal editors and reviewers (see Nicolaisen/Frandsen, 2019 on the "uncitedness" of publications).

Both frames are epistemic frames of reference because they refer to the content of the production of scientific knowledge. The main difference between them is that the first frame of reference links the content of the knowledge claim to the existing knowledge by asking whether it is new and meets the standards of production, while the second links the knowledge claim to the production of other researchers by asking whether it is useful for the further production of knowledge.

The two frames of reference also have in common their close links to social frames of reference in which the social position of researchers in their scientific community is constructed. The construction of success or failure in epistemic frames of reference simultaneously addresses the role performance of researchers as community members. Success or failure to close a gap in the

2 In FEM, we exclude from our consideration all those cases in which success or failure is also, or even primarily, judged by observers external to the scientific community.

community's knowledge affects the opportunity of researchers to maintain their identity as community members who participate in the latter's collective production of scientific knowledge. They are also closely linked to reputational success or failure because the recognition of researchers as competent members of the scientific community ultimately rests on the perception that they contribute to its knowledge production. The literature discussed in section 2 points out that the failure to make a relevant contribution also occurs when the publication of a knowledge claim is prevented, when others are credited with the knowledge claim in priority disputes, or when a published knowledge claim is ignored. The community may form the opinion that a gap in its knowledge was already closed by another member or might not perceive that a contribution was made, which may happen either because some of its members suppressed its publication or because the contribution is overlooked.

Researchers and their communities can agree or disagree in their assessments of the success or failure of research processes. For example, the case of Joseph Weber's claims to have detected gravitational waves in the 1960s and 1970s (Collins, 1981; 1999) illustrates a situation in which the researcher himself and the community diverged in their respective judgments regarding the *closure of a knowledge gap*. Weber had published several reports of observations of gravitational waves, whose existence was predicted by Einstein's general theory of relativity. Prior to Weber's publications, available measurement devices were deemed not sufficiently sensitive to be able to detect the weak signals expected from gravitational radiation. While Weber was convinced that he had been able to measure the oscillations of cosmic gravitational waves with the apparatus he built for this purpose, other members of the scientific community did not accept the validity of his claims. They held that Weber's apparatus was not sensitive enough to measure gravitational waves, and that the signal strength measured by Weber far exceeded any theoretically derived expectations. Thus, while the community did not doubt the existence of gravitational waves and agreed on the knowledge gap concerning their empirical observation, the technical implementation for their measurement was not considered viable by Weber's opponents. Their view was soon accepted in the community, and Weber's research was marginalised. This case illustrates diverging judgments in the first frame of reference, namely judgements on the validity of a knowledge claim.

A second example describes a situation where both researchers and the community concur in their assessment of failure to close a gap in the community's knowledge. For a long time, Max Delbrück searched for a complementarity principle in biology but failed to do so both in his own view and in the view of his community (Fischer/Lipson, 1988). The idea of a complementarity principle was developed by Niels Bohr for the field of quantum mechanics. Bohr

found that light can be observed either as a wave or as a particle – depending on the experimental setup – but not as both at the same time. According to Bohr, these results would have to be considered as complementary descriptions of the same truth rather than as mutually exclusive descriptions. Delbrück wanted to formulate such a principle for biology and set out to find biological phenomena which, if analysed closely enough, would lead to a paradoxical observational situation that could only be resolved by similarly complementary descriptions. A promising candidate phenomenon for this was the reproduction of life, which according to Delbrück would have to be analysed from complementary perspectives rather than with biochemical approaches that sought to break up this phenomenon into its fundamental physical components. Unfortunately for Delbrück, however, it was exactly this biochemical solution that Watson and Crick (1953) presented when they described the structure of the DNA, thereby closing the gap in knowledge in the eyes of both Delbrück and the broader scientific community. Delbrück then searched for other promising phenomena that could yield his sought-after complementarity principle in biology. However, his search remained unsuccessful, and no one else has been able to demonstrate a complementarity principle in biology.

In both cases, the contradictory respectively coinciding constructions of failure by researchers and their scientific communities concerned the question whether a relevant gap in the community's knowledge had indeed been closed. The controversy about Weber's findings arose precisely because Weber and other community members agreed that his contribution was relevant to the community's knowledge production. In other cases, however, assessments of relevance differ, and the assessment whether a knowledge gap was closed is not even made. This is the reason why many published contributions are not used by the scientific community. The large proportion of published findings in the literature which remain uncited indicate that many situations in the communal production of scientific knowledge are situations in which knowledge gaps are constructed and closed (or not closed) that are not considered relevant by the scientific community.

Since a scientific community is not able to act in the sociological sense (i.e. it cannot be considered in analogy to an individual actor), it is important to identify the mechanisms by which failure is constructed *in* (rather than by) scientific communities. A researcher's failure to close a knowledge gap or to produce a relevant contribution can be constructed in the publication process when editors or reviewers decide that a manuscript should not be published. This process, which redirects manuscripts towards minor publication channels or prevents their publication altogether, is common in all scientific communities. The scientific community delegates the decision about the success or failure of particular knowledge claims to some of its members who act as

editors and reviewers. Once knowledge claims are published, any member of a scientific community may challenge them and may even demand their retraction, a mechanism that has gained some prominence in the last two decades. In some cases, such challenges lead to a controversy about knowledge claims, which may also end with their rejection (as has happened with Weber's claim to have observed gravitational waves). The main mechanism by which the failure to contribute a relevant knowledge claim is constructed, however, is the scientific community ignoring knowledge claims. The collective 'decision' of a scientific community that a knowledge claim is not a relevant contribution is the aggregate effect of community members not using it in their subsequent research. Since this disregard by scientific communities does not discriminate between knowledge gaps that were and were not closed, we decided to exclude it from our study.

However, a researcher's failure to close a particular knowledge gap may nevertheless constitute a relevant contribution to the community's knowledge if the failure and knowledge about its causes affect the further production of knowledge. This argument is often brought forward when the lack of publications on negative results is lamented. Reporting failed research can advance the knowledge production of the scientific community by informing other researchers about the limitations of approaches or about scientifically interesting reasons for failure. In these cases, the failure to close a gap in the community's knowledge constitutes a relevant contribution. Again, researchers' constructions on that issue may differ from those of their colleagues.

These theoretical distinctions show that significant epistemic failure can be constructed in two frames of reference – as the failure to close a knowledge gap or as the failure to make a relevant contribution – and can be constructed either by researchers or by their scientific communities. These two distinctions already point to different forms of failure even before differences between disciplines are considered.

4. Methodological Considerations: How to Empirically Identify Likely Cases of Significant Epistemic Failure

There are good reasons why the systematic empirical investigation of significant epistemic failure should fail. Some specific forms of failure may remain invisible. The failure to formulate a research question – to identify a relevant gap in the community's knowledge – does not lead to projects being designed or funded and thus remains largely invisible to both scientific communities

and external observers. Failed research projects may not lead to publications because researchers don't know what to publish. If they lead to publications, these publications may remain neglected by the community and thus barely visible. Successful projects may not lead to publications because they are constructed as failed by representatives of the scientific community (editors and reviewers), and will thus stay invisible. Finally, the large number of publications that is ignored by scientific communities can be considered as failing to make a relevant contribution. As mentioned before, this form of failure is difficult to delineate because ignorance is not easy to operationalise.

In this section, we discuss the methodological challenges involved in the empirical identification of potential cases of failure through the application of unobtrusive methods. In our project on forms and effects of failure, we plan to investigate cases of significant epistemic failure by interviewing researchers involved in failed projects. To be able to recruit researchers, we first need to identify likely cases of such failure. As already mentioned in the introduction, this is not easy because failure often remains invisible in publications.

We are currently pursuing three main strategies to search for cases of significant epistemic failure. First, we assume that funding programmes for '*high-risk, high-reward' research* do indeed fund risky projects, and that some of these risky projects failed. We therefore look for projects that are funded by such programmes and yielded no or few publications (4.1). Second, we search for *publications that were retracted* due to 'honest errors' in a retraction database (4.2). Finally, among *publications of negative results* we seek to identify cases in which negative results indicate failure because researchers' expectations of positive results were disappointed (4.3). These strategies have the potential to capture some but not all variations of visible failure. Each strategy is designed to be unobtrusive, which is why we focus on publications as an (admittedly crude) indicator of knowledge claims being produced and set thresholds on time spans as well as publication and citation frequencies. Since our goal is to identify promising candidates for failure, each of the strategies is designed to favour precision over recall. It does not matter too much for our investigation that we are unlikely to identify all cases of failure as long as the cases we do identify are likely to be instances of significant epistemic failure.

We limit our search for cases of significant failure to mathematics, the natural sciences and the life sciences in order to reduce the variance of possible forms and effects of failure. While significant failure does occur in the humanities and social sciences, it takes on specific forms, which are often difficult to identify. Failed projects are more likely to still yield publications, and publications representing significant failure are more likely to be ignored rather than

retracted or contradicted. We also excluded engineering and computer science because of the likelihood of external actors being involved in the construction of success or failure of research processes. In these cases, failure might be constructed because of non-epistemic concerns.

4.1 *'High-risk, high-reward' Research Funding Programmes*

In response to the perception that traditional funding programmes tend to be risk-averse and are therefore ill suited to support scientific breakthroughs (Heinze, 2008; Wang et al., 2017) funding agencies have implemented funding programmes that specifically encourage research proposals with higher inherent risks. If the projects funded by these programmes are indeed more risky than conventional projects, the likelihood that they fail should also be higher. Therefore, projects funded by 'high-risk, high-reward' programmes provide a promising source to identify failed research processes.

We search for projects yielding few or no project-related publications because this indicates that either the researchers themselves or those deciding on submitted manuscripts came to the conclusion that the projects produced little that is worth publishing. This search strategy provides access to significant failure that was constructed as a lack of publishable results.

One of the largest 'high-risk, high-reward' funding programmes is that of the European Research Council (ERC). According to the ERC, its "founding principle (…) is to target frontier research by encouraging 'high-risk, high-reward' proposals that may revolutionise science and potentially lead to innovation if successful" (Antonoyiannakis et al., 2009, 805–806). Many countries have established similar national funding programmes for 'high-risk, high-reward' research prior to or after the institutionalisation of the ERC. These include those of the National Institutes of Health (NIH) in the US, which provide funding opportunities in four funding programmes similar to those of the ERC under the umbrella of the "NIH Director's Awards". There are four different awards, which support individual scientists at varying career stages, and in one case teams. Other similar national funding programmes include the START programme of the Austrian Research Council, the *Experiment!* programme of the German *Volkswagen* foundation (see the contribution by Simon in this volume), and the *Reinhart-Koselleck* programme of the German Research Council (DFG). We plan to include several of these other programmes if data availability permits.

In contrast to the other funding agencies, both ERC and NIH provide lists of funded projects and their publications on their websites. The selection of candidate projects for significant failure included the following steps (Figure 11.1). In the case of the ERC projects, we had to filter out projects from the humanities, engineering, computer science and the social sciences, which are not

included in our investigation. In the case of the NIH, this step was not necessary because all awards fund biomedical or public health research. Second, we limited the possible candidates to research projects that should officially end no later than 2019, as we assumed that projects ending later might *not yet* have fully published their results. Overall, this resulted in 3689 possible candidates within the ERC programme (1831 starting grants, 473 consolidator grants, 1328 advanced grants and 57 synergy grants) and 678 possible candidates within the NIH programme (143 NIH Director's Pioneer Awards, 56 NIH Director's Early Independence Awards, 360 NIH Director's New Innovator Awards and 119 NIH Director's Transformative Research Awards).

Our next step was to look into databases listing these projects and to identify projects with no more than three publications. This is an arbitrary threshold that is likely to produce false positives (successful projects that had few publications) and false negatives (failed projects with more than three publications). However, setting such a threshold appears to be the only way to identify candidates for failed projects unobtrusively (without involving constructions by principal investigators or other members of their scientific community).

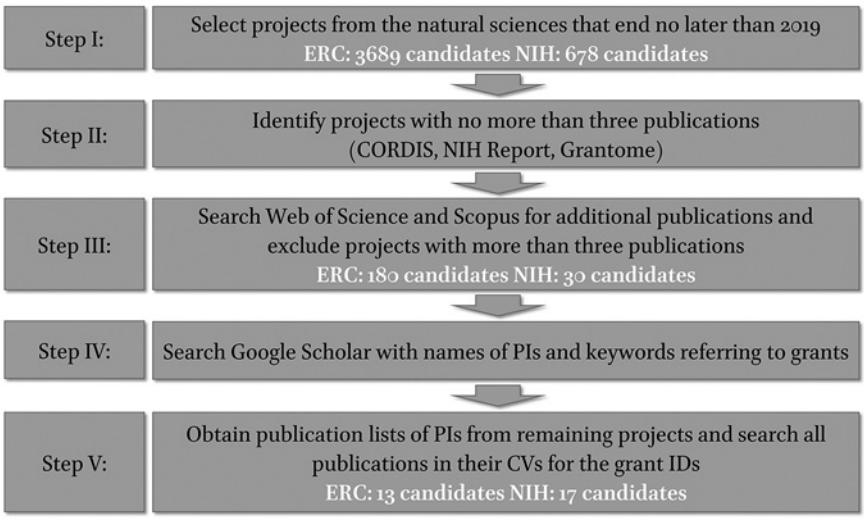

Figure 11.1 Search strategy for candidate projects from ERC and NIH funding programmes

In the case of the ERC, we could use CORDIS, the official project database of the European Commission which lists results of all research projects funded by the EU. For NIH projects, we used the databases RePORT and Grantome.[3]

3 The websites can be found at the following links: https://cordis.europa.eu/ (CORDIS), https://report.nih.gov/ (RePORT) and https://grantome.com/ (Grantome).

Since these databases do not have complete records of project publications, our next step was to search additional databases (Web of Science and Scopus). We searched with the Grant IDs of our candidates in both databases. Not all Grant IDs were unique, so we had to verify whether the identified publications belonged to the NIH or the ERC. Since some publications can be found in more than one database, we had to compare and compile the results from the different databases and make sure that the remaining research projects had three or less unique publications.

With these two steps, we were able to significantly reduce our list of possible candidates, but we decided that we still had to further verify our results, as these databases could still have missed project-related publications. Therefore, our next step involved a search for publications in Google Scholar. Our strategy was to search for the principal investigators of the research projects and combine their names with keywords like, e.g. the Grant IDs, the acronyms of the projects, "European Research Council", "NIH" or the name of the respective grants (e.g. "starting grant" or "pioneer award"). As Google Scholar searches for keywords in the full texts of articles, we were able to further reduce our sample significantly with this strategy.

As a final step, we decided to obtain the publication lists of all the remaining principal investigators and to screen the "acknowledgement" or "funding" sections of all the articles they had published after the respective research projects had started. We looked for any statements that indicated a funding by the research project. This manual search technique was an effective addition to the database-related semi-automatic search techniques explained above. The final result includes 30 promising candidates for failed projects of which 13 were ERC research projects (seven starting grants and six advanced grants) and 17 NIH research projects (five NIH Director's Pioneer Awards, two NIH Director's Early Independence Awards, five NIH Director's New Innovator Awards and five NIH Director's Transformative Research Awards).

4.2 *Retractions of Research Articles*

Retracting a scientific publication means removing it from a journal. Journal articles can be retracted for many different reasons, the most frequent of them being (intentional) misconduct and (unintentional) 'honest errors'. A study on retractions in *Science* between 1983 and 2017 shows that 48% of the retractions were due to "unintentional errors" and 33% due to "intentional errors", while for 19% the intention was unclear (Andersen/Wray, 2019; see also Vuong, 2020 for problems with retraction notices). In the context of our research project, we are only interested in retractions due to honest errors, especially those that affected the main conclusions of a study, e.g. because results are not

reproducible or invalid. In the case of these retractions, the authors expected to fill a gap in the literature and to make a relevant contribution but ultimately were not able to do so. A retraction reveals these failures caused by honest errors and makes them accessible to our research.

For our empirical search, we relied on the website Retraction Watch and the associated database.[4] Retraction Watch was founded in 2010 by Ivan Oransky and Adam Marcus because they realised that retractions and their reasons were rarely publicised. The database contains more than 25.000 entries, with some of the retracted articles dating back to the 1920s. It enables a search for retractions by their reasons, codes for which were added by the database owners.[5] Coded reasons fall into three broad categories, namely misconduct (e.g. plagiarism, falsification of data), formal problems (e.g. lack of approval from author), and errors. As we are trying to identify retractions that indicate significant epistemic failure, we searched for 'honest errors' in research, i.e. codes that refer to the content of the published article and do not signal misconduct. Specifically, these were entries with the codes "error in results and/or conclusions" or "results not reproducible". Both codes imply a failure in the research process that has led to unpublishable results.

However, retractions usually have multiple codes assigned to them and we are not interested in retractions due to fraud or ethical violations. In order to select only cases of retractions due to honest errors, we had to filter out publications with additional codes indicating misconduct of any kind. We also excluded retractions exclusively categorised as social science, humanities, or computer science and engineering. Finally, we decided to focus on retractions between 2010 and 2021, because the Retraction Watch website started to provide further information on some cases occurring after its implementation in 2010 but not on older cases. This additional information (in the form of blog entries) was necessary to unambiguously identify the reasons and circumstances of retractions. The elimination of retractions between 2010 and 2021 for which no further information was available on the Retraction Watch website left 161 entries for "error in results and/or conclusions" and 101 entries for "results not reproducible".

4 See: https://retractionwatch.com/ and http://retractiondatabase.org.

5 The database owners collect information about retracted articles. There is no unambiguously described procedure but apparently a search of journals and submissions by researchers are the main sources for new records. As with the other strategies, the search among retractions is aimed at high precision rather than perfect recall, which is why omissions in the database do not constitute a problem. For the codes used by the database owners, see: https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-b-reasons/.

In the next step, we read the blog entries and, if necessary, the retraction notices of all 262 remaining retractions. It became evident that our sample still included a few clear cases of misconduct, while for other cases misconduct could not be ruled out completely. This happened, for example, in cases where one of the authors had multiple retractions that were categorised differently, with some of the retractions referring to fraud and others remaining ambiguous. We retained only cases in which there was no indication of misconduct or of formal issues as reasons for retraction.

Further exclusions included a case in which the article was riddled with errors and mistakes, a few duplicates, the retraction of a conference abstract and all cases in which no clear reasons for the retraction could be obtained from the database or blog entries. We also decided to filter out controversial retractions, i.e. cases in which one or more authors disagreed with the retraction. As outsiders, we were not able to assess whether the retraction was justified or whether (honest) error had occurred. Therefore, we eliminated these controversial cases.

We then selected articles with substantial problems, i.e. articles whose authors were not able to correct their mistakes without making the article and its results irrelevant. In most cases the severity of problems could already be derived from context information on the websites. The best indicator, however, was the inability of authors to correct their mistakes and to publish the results again in a different publication.

Our last step was the identification of retracted publications that were visible to the scientific community. We defined 'being visible' to the scientific community as receiving at least 30 citations in the Web of Science. We are aware that numerical limits are always arbitrary but it is unlikely that all 30 citations are perfunctory, which means that the cited publications were not ignored by the community. We did not distinguish between citations prior to or after the retraction because retracted publications continue to be cited, and citation context analyses demonstrate that the retraction is rarely mentioned by citing publications (Teixeira da Silva/Bornemann-Cimenti, 2017; Schneider et al., 2020).

The result includes nine retractions in the category "results not reproducible" and 14 retractions in the category "error in results and/or conclusions". These 23 cases represent all identifiable non-controversial retractions between 2010 and 2021, which were due to grave honest errors that made results unpublishable, and which were not ignored by the scientific community. We consider these articles as good candidates for significant epistemic failure.

4.3 *Negative Results*

Publications of negative results are the third main starting point of our search for cases of significant epistemic failure. Since the early 2000s, several tools have been created with the aim to increase the visibility of negative findings, among them:

- journals specialising in the publication of negative results;
- sections in 'regular' journals dedicated to the publication of negative results;
- special issues of 'regular' journals dedicated to the publication of negative results;
- dedicated online collections of articles publishing negative results;[6] and
- search options offered by 'regular' journals that allow to filter published articles for negative results.[7]

To identify publication channels for negative results, online searches with the Google search engine were conducted. We first used the search strings "journals publishing negative results", "special issue negative results", and "article collection negative results". Additional search strings were then developed to complement the initial findings.[8] Finally, Ulrich's Periodicals Directory[9] was searched, using the terms "negative", "null", "insignificant", "non significant", "inconclusive", "non conclusive", "irreproducible", "non reproducible", "fail*", and "unsuccessful". The search led to some publication channels for negative

6 See, e.g. the collections by *Nature: Scientific Reports* (https://www.nature.com/collections/gcifjebabg/) or *PlosOne* (https://collections.plos.org/collection/missing-pieces/).

7 Journals "considering" negative results, such as *PeerJ* or *ACS Omega*, were not taken into account in our search process, because they provided no obvious way to identify negative results, e.g. via keyword-search options.

8 These included the search strings ("negative result*" OR "negative finding*" OR "negative data"), ("null result*" OR "null finding*" OR "null effect*" OR "null hypothesis"), ("insignificant result*" OR "insignificant finding*" OR "insignificant effect*"), ("non significant result*" OR "non significant finding*" OR "non significant effect*"), ("inconclusive result*" OR "inconclusive finding*" OR "inconclusive effect*"), ("non conclusive result*" OR "non conclusive finding*" OR "non conclusive effect*"), ("non reproducible result*" OR "non reproducible finding*" OR "non reproducible experiment*" OR "non reproducible research"), ("irreproducible result*" OR "irreproducible finding*" OR "irreproducible experiment*" OR "irreproducible research"), ("failed experiment*" OR "failed research" OR "failed project*" OR "experimental failure*" OR "research failure*" OR "project failure*" OR "scientific failure*" OR "failure in science" OR "failures in science"), and ("unsuccessful experiment*" OR "unsuccessful research" OR "unsuccessful project*") which were coupled via an AND-operator with the search strings (journal AND publish*), ("journal section" OR "thematic section" OR "theme section"), or ("special issue" OR "special topic" OR "special section"), respectively.

9 Ulrich's Periodicals Directory (https://www.ulrichsweb.com/) is a comprehensive bibliographic database covering academic and non-academic serials, including titles that have ceased to publish.

results directly. It also returned a range of documents including blog posts, commentaries, opinion pieces and similar material as well as journal papers discussing the publication of negative results, providing (partial) lists of dedicated publication channels (e.g. Teixeira da Silva, 2015; Sayao et al., 2021). The search also pointed to some journals publishing negative results that already ceased to exist and are now only accessible via the Wayback Machine – Internet Archive.[10] For some journals, no active websites could be found. The publication channels in the consolidated list were then used to retrieve all the accessible published material. From this material, only original research articles were included in the final collection of papers to be searched for candidates for significant failure.[11]

In total, 76 publication channels for negative results were identified. Of these, 35 are journals dedicated to the publication of negative results, 20 are special issues, 21 are journal sections, article collections, or journals offering options to search for negative results. Only 20 of the 35 dedicated journals have ever published an original research article. Of these, five journals have only published one article ever. In eight cases, the publication activity of the journal could not be determined.

A total of 1.171 research articles to be taken into further consideration was retrieved, of which 118 have appeared in special issues. While the journals cover all disciplines, the distribution across fields is very skewed. The vast majority of articles has been published in biomedical and pharmaceutical journals, and to some extent in biology and ecology. Psychology is also well represented, while the physical sciences, social sciences, humanities and engineering disciplines are largely absent from the sample. Figure 11.2 shows the number of articles published in each year between 1 January 2000 and 10 September 2021. Overall, the prevalence of papers published across various specialties appears to be in accordance with the prominence of psychology and health research in the discourse about a publication bias against negative results, which has first emerged in these disciplines (Marks-Anglin/Chen, 2020).

10 See: https://archive.org/.

11 Two important publication channels for negative results, *BMC Research Notes* and *Neurobiology of Aging*, publish negative results in the form of research notes and short communications. In these cases, these were also included. They comprise 126 articles which in the following are treated as research articles.

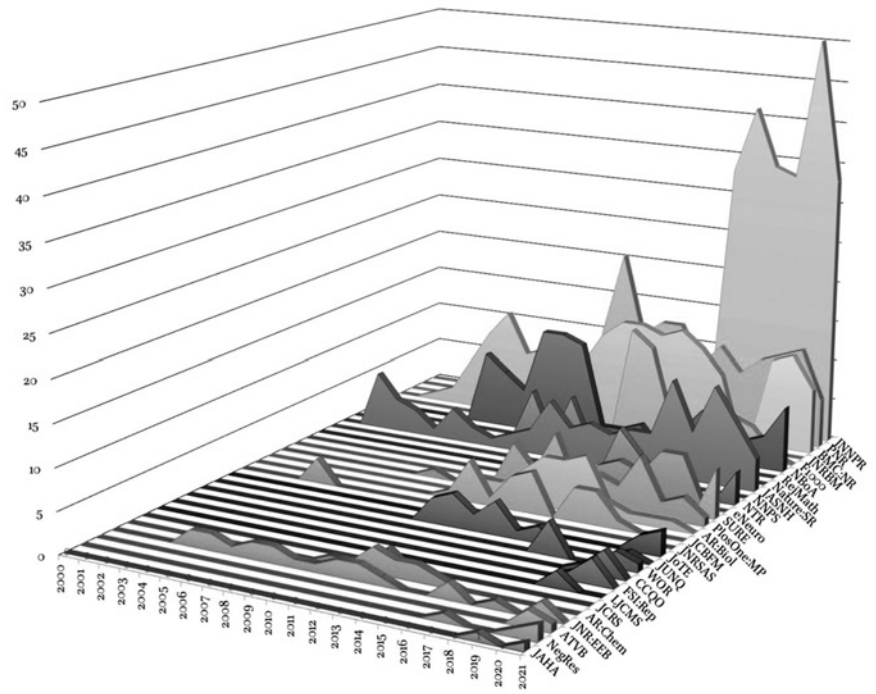

Figure 11.2 Number of papers published per year in publication channels dedicated to negative results[12]

12 Abbreviations of publication channels on the Z-axis, in the order of their listing from front to back: JAHA = Journal of the American Heart Association; NegRes = Negative Results; ATVB = Arteriosclerosis, Thrombosis, and Vascular Biology; JNR:EEB = Journal of Negative Results – Ecology and Evolutionary Biology; AR:Chem = The All Results Journals: Chem; JCRS = Journal of Contradicting Results in Science; LJCMS = Living Journal of Computational Molecular Science; FSI:Rep = Forensic Science International: Reports; CCQO = Circulation: Cardiovascular Quality and Outcomes; WOR = Wellcome Open Research; JUNQ = Journal of Unsolved Questions; JoTE = Journal of Trial and Error; JNRSAS = Journal of Negative Results in Speech and Audio Sciences; JCBFM = Journal of Cerebral Blood Flow and Metabolism; AR:Biol = The All Results Journals: Biol; PlosOne:MP = PlosOne – The Missing Pieces; SURE = Series of Unsurprising Results in Economics; eNeuro = eNeuro; NTR = Neurotrauma Reports; NNPS = New Negatives in Plant Science; JASNH = Journal of Articles in Support of the Null Hypothesis; Nature:SR = Scientific Reports – Negative Results; RejMath = Rejecta Mathematica; NBoA = Neurobiology of Aging; F1000 = F1000 Research; JNRBM = Journal of Negative Results in Biomedicine; BMC:NR = BMC Research Notes – Negative Results; PNR = Journal of Pharmaceutical Negative Results; JNNPR = Journal of Negative & No Positive Results. Special issues of regular journals are excluded from the figure, as well as publication channels that have not published more than one article (these include five dedicated journals, one journal section, and six journals providing search filters).

As can be seen in Figure 11.2, only few publication channels do continuously publish research articles. General publication activity is very low, and only five publication channels cover 61% of the research articles included in our sample. The majority of journals thus seems to either suffer from a lack of submissions, or submitted manuscripts are not of high enough quality. Most journals depicted in the figure are barely visible and are not indexed in major databases such as the Web of Science. Even among the five most productive journals (which contribute the majority of articles in our sample), only NBoA (*Neurobiology of Aging*) is currently indexed in the Web of Science.

The overall picture of publication channels dedicated to negative results confirms the widely held impression of a tendency to devaluate negative results in the publication practice of scientific communities. This perception is further reinforced by the low total number of articles with negative results in comparison to the overall scholarly record of published papers.

To identify candidates for significant failure, we are currently conducting a content analysis of the abstracts of articles publishing negative results. The content analysis shows that different types of findings are published under the label 'negative results', and that not all of them point to significant epistemic failure as defined by us. The negative results which are reported in our sample typically refer to findings that do not support the claim that an effect exists or the claim that a method functions as assumed. According to our definition, findings of this kind only indicate failure if the goal of a research process was to demonstrate the existence of an effect or the functioning of a method. If the goal of a research process was to prove that an effect does not exist or that a method does not work, the negative results constitute success rather than failure. Similarly, research that was exploratory, i.e. open with regard to the direction of results, must be considered successful if the results are negative. The following abstract of a paper illustrates this case:

> The translocator protein, a microglial-expressed marker of neuroinflammation, has been implicated in Alzheimer's disease, which is characterized by alterations in vascular and inflammatory states. A TSPO variant, rs6971, determines binding affinity of exogenous radioligands in vivo; however, *the effect of these altered binding characteristics on inflammatory and cerebrovascular biomarkers has not been assessed*. In 2345 living subjects (Alzheimer's Disease Neuroimaging Initiative, n = 1330) and postmortem brain samples (Religious Orders Study and Memory and Aging Project, n = 1015), we analyzed effects of rs6971 on white matter hyperintensisites, cerebral infarcts, circulating inflammatory biomarkers, amyloid angiopathy, and microglial activation. We found that rs6971 does not alter translocator protein in a way that impacts cerebrovascular and inflammatory states known to be affected in dementia. (Felsky et al., 2016, abstract, emphasis added)

The abstract does not indicate that the research was guided by the (disappointed) expectation that an effect exists. The only possible failure of this kind of exploratory research would be the failure to answer the initial question ('what is the effect of rs6971's altered binding characteristics on certain types of biomarkers?') either positively or negatively because inconclusive or otherwise uninterpretable results were obtained.

We exclude reports of research processes aimed at producing negative results, reports of open exploratory research processes, and ambiguous cases. Instead, we search for studies reporting that specified expectations regarding study outcomes were disappointed. These expectations can be differentiated in two dimensions. With regard to the sources of expectations, we can distinguish between research that is based on the authors' own expectations and research based on expectations of other community members. In a second dimension, we distinguish between empirical/theoretical and methodological research goals. These two distinctions led to a coding scheme that is currently applied to the 1.171 articles obtained from the different publication channels for negative results (Table 11.1).

		Type of research goal	
		Empirical/theoretical	Methodological
Source of expectations	Community	Community hypothesis	Evaluation of methods
	Researchers	Hypothesis of researchers	Development of methods

Table 11.1 Coding scheme for negative results

We now briefly illustrate the four cases with examples already obtained in the content analysis. First, there is the case of a disappointed community expectation regarding empirical/theoretical research goals ("Community hypothesis"). In this case, generalised community expectations regarding the existence of a particular effect or causal relationship guide a research process, which, however, leads to a negative result. The negative result calls these expectations into question:

> Low-frequency magnetic fields (LF-MF) generated by power lines (…) are classified as possibly carcinogenic by the World Health Organization. *Epidemiological studies indicate that LF-MF might propagate neurodegenerative diseases like Alzheimer's disease* (*AD*) *or amyotrophic lateral sclerosis* (*ALS*). We conducted a comprehensive analysis to determine whether long-term exposure to LF-MF (…) interferes with disease development in established mouse models (…). [Our

> results suggest] that *LF-MF do not affect cellular processes involved in the pathogenesis of AD or ALS*. (Liebl et al., 2015, abstract, emphases added)

"Evaluation of methods" represents cases in which methodological expectations of the community are contradicted by negative results. This happens when methods which are established in the community are used but in a particular case of application did not produce the expected results:

> Fluorescently labelled nanoparticles are *routinely used in Correlative Light Electron Microscopy* (CLEM) (…). Herein we show (…) that (…) the observed fluorescent signal in fact arises from a large population of untagged fluorophores; rendering these labels *potentially ineffective and misleading to the field*. (Miles et al., 2017, abstract, emphases added)

If expectations were formulated by the researchers themselves, the success or failure of research processes is completely their own ("Hypothesis of researchers"). The following abstract gives an example of disappointed empirical/theoretical expectations:

> Medication effect is the sum of its drug, placebo, and drug*placebo interaction effects. *It is conceivable that the interaction effect involves modulating drug bioavailability*; it was previously observed that being aware of caffeine ingestion may prolong caffeine plasma half-life. This study was set to evaluate such concept using different drugs. (…) *This study couldn't confirm that awareness of drug ingestion modulates its bioavailability*. (Hammami et al., 2017, abstract, emphases added)

Finally, methodological expectations of the researchers can be disappointed when they develop new methods that do not work as expected ("Development of methods"). The major difference to the evaluation of already existing methods is that these methods are new, and the community has not yet developed expectations concerning their validity and reliability:

> *To find a practical clinical tool to assess DIP* in patients with severe mental illness (SMI), the association between blink rate and drug-induced parkinsonism (DIP) was assessed. (…) There is a significant association between blink rate and DIP as diagnosed on the UPDRS. However, *blink rate sensitivity and specificity with regard to DIP are too low to replace clinical rating scales in routine psychiatric practice*. (Mentzel et al., 2017, abstract, emphases added)

Based on the current stage of our coding process, we expect a significant reduction of the material resulting from the exclusion of studies that do not clearly indicate initial expectations of the outcome of their research in their abstracts.

In addition, the distinctions leading to the table and the four cases of disappointed expectations raise an interesting question for later interview-based

research. If a researcher conducts an investigation that is guided by, and thus tests, expectations of their scientific community: Whose failure is constructed in what ways if the goals of the research process are not achieved?

4.4 *Comparison of the Search Strategies*

The three search strategies operationalise our definition of significant epistemic failure in different ways and therefore capture different forms of failure.[13]

The search strategies differ in their *unit of analysis*. We defined significant epistemic failure at the level of research processes or research programmes in order to distinguish it from failure in everyday research practice, which can be fixed without a substantial modification of research goals. While the search for failed research projects is likely to lead to the identification of failed research processes, the search for retracted publications and publications of negative results targets a smaller unit of analysis. Single publications can point to failures at *any* level of a research process. Since most research processes lead to more than one publication, a retracted publication or a publication containing negative results does not by itself indicate the failure of a whole research process. Therefore, projects with few or no publications offer a more direct access to significant failure than retractions and published negative results, which may only do so in some cases.

Determining the role of single retracted publications or publications with negative results in research processes requires additional analytical steps. If funding information is available for the publication, they can be associated with projects, and other publications belonging to the same project can be checked for indications of failure. Additional information about the context of single publications can also be derived from bibliometric data. We will identify authors of a retracted publication or publication with negative results and analyse their publications in a period beginning five years before and ending two years after the identified publication. A lower number of publications in this period could indicate a failed project, as it would point to a lack of publishable findings in this period. However, this indicator may be blurred when an author takes part in multiple projects in parallel so that the failure of a single project may become 'overshadowed' by the publication output stemming from other, parallel, projects.

A second, more sophisticated, use of bibliometric data to determine the relation of a publication to broader project-related research goals would be the reconstruction of "research trails" (Gläser/Laudel, 2015a). Research trails are used to identify the research topics defining the cognitive career (Gläser/

13 So far, we have not yet encountered any overlaps in the form of failed projects being associated with either retractions or the publication of negative results.

Laudel, 2015b) of a researcher. They are produced by thematically clustering the oeuvre of a researcher and displaying the clusters in their chronological order. This approach makes it possible to determine the position of retracted papers and papers reporting negative results in relation to the topics worked on by an author. Thus, if a paper reporting negative results or a retracted paper is the last one published on a particular topic this might indicate that the research goals relating to that topic were not achieved, and the topic was abandoned. Conversely, if a negative result or a retracted paper has appeared relatively early in the publication history on a given topic, this might indicate that the goals of these research processes were achieved, and the failure was not significant. Finally, publications with negative results concerning a community expectation might be assessed by citation counts because a citation count of the paper would indicate the community's response to the refutation of its expectations. If a negative result indicating a failure on the part of the community is not cited, then it is also highly unlikely that such a construction of a community failure has indeed been successful.

The search strategies also differ in the *amount of* '*noise*' in the data they utilise. In the search for failed research projects, we can draw on quite robust data from project and publication databases even though there is still considerable effort required to compensate for the incompleteness of most publicly available databases. In contrast, retractions and negative results can be produced by other processes than significant failure. Retractions in the database that are labelled as 'honest errors' may be coupled to other reasons for retraction such as misconduct or formal issues. Additional information is necessary for the identification of promising candidates for significant failure. Negative results do not necessarily point to the failure of researchers to achieve their goals, i.e. to disappointed expectations of the authors of a study, or of community members. Therefore, it was necessary to develop a coding scheme to identify promising cases.

Analysing textual data about the *content* of research which is necessary especially when abstracts of published negative results are coded and to a lesser extent for the filtering of retracted papers, faces particular difficulties. Interpreting information about research content is notoriously difficult for outsiders due to the esoteric nature of specialised scientific knowledge production. Counting publications of risky projects does not require familiarity with esoteric knowledge and thus renders this search strategy more 'transparent' – although it is not necessarily more valid.

A second concern relating to the textual information at hand has to do with the incompleteness of information on the research process, especially as contained in the abstracts and full texts of publications on negative results. While

abstracts are likely to reflect the major *findings* of a published study, we cannot as easily assume that abstracts or full texts contain faithful information on the original goals of a research process and the *expectations* underlying it (Knorr-Cetina, 1981). The ex-post rationalisations that are an inextricable aspect of the genre of the experimental article are an additional hurdle for the reconstruction of research goals and expectations of its authors.

5. Conclusions

Although the theoretical considerations and empirical search strategies presented in this chapter document only first steps in an ongoing project, three conclusions can be drawn from our work so far. First, contrary to the common perception of scientific failure as an experience of individual researchers or research groups, significant epistemic failure always involves a complex relationship between researchers and their scientific communities. This is of course true in the trivial sense that all research processes use the community's knowledge and thereby test it. If researchers fail, their application of the community's knowledge fails. However, the scientific community is also involved beyond this fundamental relationship. Our preliminary analysis of articles publishing negative results demonstrated that the extent to which researchers test hypotheses of the scientific community rather than hypotheses of their own varies. Research can be conducted based on own expectations, on expectations of colleagues, or on expectations that are widely held in a scientific community. If this research fails, it may reflect a failure of the scientific community.

The involvement of the community in the construction of failure varies, too. If researchers decide that no knowledge worth publishing emerged from their research process, this individual decision is still made by pitching the outcomes of a research process against the community's knowledge. The role of the community becomes stronger when researchers deem their findings worthy of publication but anticipate that their community will hold a different opinion and therefore don't try to publish. If they do try to publish and community members who serve as editors or reviewers foil their attempt, the community influence becomes even stronger. Finally, the community may implicitly declare a published contribution as irrelevant or failed by not using it in its further work. Thus, the knowledge gap, the knowledge claim closing it and the assessment of success or failure are jointly constructed by researchers and their communities, and the influence of participants in this joint construction process varies.

Second, we analytically distinguished several ways in which the failure of research processes or research programmes may be constructed. Forms of failure differ in their visibility to outside observers, which creates specific challenges to their empirical identification. Failure that does not leave traces in publications can be identified by ethnographic observations, interviews, surveys, or by comparing stated goals of research processes to publication outputs. Since significant epistemic failure appears to be rare and unpredictable, its empirical identification via ethnographies and interviews is likely to fail. Ethnographies and interviews are limited to the investigation of few cases and thus cannot be used to screen a large number of research processes for instances of failure. Surveys are more likely to be successful because they cast a wider net but would need a careful operationalisation of the concept of failure in order to align the research interest with respondents' everyday understanding of failure.

If research processes take the form of funded projects, their stated goals can be compared to publication outputs. We apply this strategy but limit it to funding programmes for 'high-risk, high-reward' research because we expected a higher likelihood of failure among high-risk projects. The number of cases with few or no publications we could identify illustrates that we search for needles in a haystack, and that the extension of this strategy to all funded projects would be very inefficient. Our preliminary results suggest that researchers are intent on designing projects that are unlikely to fail, and that they are very proficient in securing success.

Another opportunity to compare goals to outcomes, which we have not yet explored, is comparing the pre-registration information about studies to their actual outcomes. The pre-registration of study goals and designs is becoming increasingly common in medical research and some other fields. It offers an opportunity to compare the actual outcomes to the registered goals, and to identify pre-registered studies whose results were never published.

Obtaining traces of significant failure from publications is difficult for another reason. A publication is likely to represent only a partial outcome of a research process, which makes it impossible to draw conclusions from a 'failed' publication (a publication that was retracted or a publication containing negative results) to a failed research process. A publication only provides an entry point for the reconstruction of oeuvres of research processes, without which it is impossible to decide whether a research process failed.

Third, we identified possible sources for disciplinary variations of significant failure. The extent to which researchers use hypotheses developed in their community rather than hypotheses they developed themselves, and the extent to which community members agree on the success or failure of research

processes is likely to depend on the degree of codification of the community's knowledge, which "refers to the consolidation of empirical knowledge into succinct and interdependent theoretical formulations" (Zuckerman/Merton, 1973 [1972], 303). A common theoretical background and strong methodological standards are likely to support the existence of shared expectations in a community as well as the existence of shared yardsticks for the assessment of success or failure. The uneven distribution of retractions and publications containing negative results, which are heavily concentrated in biomedical fields, also suggests that some causes for significant failure may be unevenly distributed between scientific fields.

How can our project fail? The preliminary results of our search indicate that we need to 'update' our introductory remark on the triangle of empirical ubiquity, epistemological admiration and practical neglect. While failures in everyday scientific work are indeed ubiquitous but neglected in both epistemological discussions and the communication processes of scientific communities, significant epistemic failure appears to be rare, epistemologically admired, and practically neglected. The rarity of significant epistemic failure appears to be caused by two kinds of behaviour. First, scientists try hard to avoid significant epistemic failure. Ethnographic studies of science have demonstrated that scientists invest much effort in 'making things work' in the laboratory, i.e. in not failing (see e.g. the studies we referred to in section 2). We should also remember that funded projects that do not lead to new knowledge pose a major risk for further funding or for career advancement. Having to retract a publication, even if the reason is an 'honest error', is regarded as similarly disadvantageous. Second, even significant epistemic failure does not appear to be of interest to the scientific community. Communicating such failure is still difficult for researchers. Publications containing significant failure are not written, not accepted, or do not find much attention. Thus, a rare but important aspect of scientific research turns out to be a very elusive phenomenon.

References

Andersen, L. E. / Wray, K. B. (2019): Detecting errors that result in retractions. In: Social Studies of Science, 49 (6), p. 942–954.

Antonoyiannakis, M. / Hemmelskamp, J. / Kafatos, F. C. (2009): The European research council takes flight. In: Cell, 136 (5), p. 805–809.

Barber, B. (1961): Resistance by Scientists to Scientific Discovery. In: Science, 134 (3479) 1961, p. 596–602.

Cadeddu, A. (2000): The Heuristic Function of 'Error' in the Scientific Methodology of Louis Pasteur: The Case of the Silkworm Diseases. In: History and Philosophy of the Life Sciences, 22 (1), p. 3–28.

Clark, A. M. / Thompson, D. R. (2013): Successful failure: good for the self and science. In: Journal of advanced nursing, 69 (10), p. 2145–2147.

Collins, H. M. (1981): Son of Seven Sexes: The Social Destruction of a Physical Phenomenon. In: Social Studies of Science, 11 (1), p. 33–62.

Collins, H. M. (1999): Tantalus and the Aliens. In: Social Studies of Science, 29 (2), p. 163–197.

Collins, H. M. (1975): The Seven Sexes: A Study in the Sociology of a Phenomenon, or the Replication of Experiments in Physics. In: Sociology, 9 (2), p. 205–224.

Cozzens, S. E. (1989): Social control an multiple discovery in science. The opiate receptor case. Albany, N.Y.

Dasgupta, P./David, P. A. (1994): Toward a new economics of science. In: Research Policy, 23 (5), p. 487–521.

Duyx, B. / Urlings, M. J. E. / Swaen, G. M. H. / Bouter, L. M. / Zeegers, M. P. (2017): Scientific citations favor positive results: a systematic review and meta-analysis. In: Journal of clinical epidemiology, 88, p. 92–101.

Dwan, K. / Altman, D. G. / Arnaiz, J. A. / Bloom, J. / Chan, A.-W. / Cronin, E. / Decullier, E. / Easterbrook, P. J. / Elm, E. von / Gamble, C. / Ghersi, D. / Ioannidis, J. P. A. / Simes, J. / Williamson, P. R. (2008): Systematic review of the empirical evidence of study publication bias and outcome reporting bias. In: PloS one, 3 (8).

Fam, D. / O'Rourke, M. (Eds.) (2020): Interdisciplinary and Transdisciplinary Failures. Lessons Learned from Cautionary Tales, New York.

Fanelli, D. (2012): Negative results are disappearing from most disciplines and countries. In: Scientometrics, 90 (3), p. 891–904.

Fanelli, D. (2013): Positive results receive more citations, but only in some disciplines. In: Scientometrics, 94 (2), p. 701–709.

Felsky, D. / Jager, P. L. de / Schneider, J. A. / Arfanakis, K. / Fleischman, D. A. / Arvanitakis, Z. / Honer, W. G. / Pouget, J. G. / Mizrahi, R. / Pollock, B. G. / Kennedy, J. L. / Bennett, D. A. / Voineskos, A. N. (2016): Cerebrovascular and microglial states are not altered by functional neuroinflammatory gene variant. In: Journal of Cerebral Blood Flow and Metabolism, 36 (4), p. 819–830.

Fischer, E. P. / Lipson, C. (1988): Thinking about science. Max Delbrück and the origins of molecular biology. New York / London.

Franco, A. / Malhotra, N./Simonovits, G. (2014): Social science. Publication bias in the social sciences: unlocking the file drawer. In: Science, 345 (6203), p. 1502–1505.

Fraser, H. / Parker, T. / Nakagawa, S. / Barnett, A. / Fidler, F. (2018): Questionable research practices in ecology and evolution. In: PloS one, 13 (7).

Gläser, J. / Laudel, G. (2015a): A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations. In: Historical Social Research / Historische Sozialforschung, 40 (3), p. 299–330.

Gläser, J. / Laudel, G. (2015b): The Three Careers of an Academic. In: Zentrum Technik und Gesellschaft discussion paper (35).

Gläser, J. (2007): The Social Orders of Research Evaluation Systems. In: The Changing Governance of the Sciences, Eds. R. Whitley; J. Gläser. Dordrecht, p. 245–266.

Gläser, J. (2006): Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung. Frankfurt/Main / New York.

Gumpenberger, C. / Gorraiz, J. / Wieland, M. / Roche, I. / Schiebel, E. / Besagni, D. / François, C. (2013): Exploring the bibliometric and semantic nature of negative results. In: Scientometrics, 95 (1), p. 277–297.

Hagstrom, W. O. (1974): Competition in Science. In: American Sociological Review, 39 (1), p. 1–18.

Hammami, M. M. / Yusuf, A. / Shire, F. S. / Hussein, R. / Al-Swayeh, R. (2017): Does the placebo effect modulate drug bioavailability? Randomized cross-over studies of three drugs. In: Journal of negative results in biomedicine, 16 (1).

Heinze, T. (2008): How to sponsor ground-breaking research: a comparison of funding schemes. In: Science and Public Policy, 35 (5), p. 302–318.

Hon, G. (1995): Going Wrong: To Make a Mistake, to Fall into an Error. In: The Review of Metaphysics, 49 (1), p. 3–20.

Hon, G. (1987): On Kepler's awareness of the problem of experimental error. In: Annals of Science, 44 (6), p. 545–591.

Hon, G. (1989): Towards a typology of experimental errors: An epistemological view. In: Studies in history and philosophy of science, 20 (4), p. 469–504.

Hook, E. B. (Ed.) (2002): Prematurity in scientific discovery. On resistance and neglect, Berkeley, Los Angeles, London.

Knorr-Cetina, K. D. (1981): The manufacture of knowledge. An essay on the constructivist and contextual nature of science. Oxford.

Kuhn, T. S. (1962): The Structure of scientific revolutions. Chicago.

Latour, B. / Woolgar, S. (1986 [1979]): Laboratory life. The construction of scientific facts. Princeton.

Liebl, M. P. / Windschmitt, J. / Besemer, A. S. / Schäfer, A.-K. / Reber, H. / Behl, C. / Clement, A. M. (2015): Low-frequency magnetic fields do not aggravate disease in mouse models of Alzheimer's disease and amyotrophic lateral sclerosis. In: Scientific reports, 5.

Loscalzo, J. (2014): A celebration of failure. In: Circulation, 129 (9), p. 953–955.

Lynch, M. (1985): Art and artifact in laboratory science. A study of shop work and shop talk in a research laboratory. London.

Marks-Anglin, A. / Chen, Y. (2020): A historical review of publication bias. In: Research synthesis methods, 11 (6), p. 725–742.

Mentzel, C. L. / Bakker, P. R. / van Os, J. / Drukker, M. / Matroos, G. E. / Tijssen, M. A. J. / van Harten, P. N. (2017): Blink rate is associated with drug-induced parkinsonism in patients with severe mental illness, but does not meet requirements to serve as a clinical test: the Curacao extrapyramidal syndromes study XIII. In: Journal of negative results in biomedicine, 16 (1).

Merton, R. K. (1957): Priorities in Scientific Discovery: A Chapter in the Sociology of Science. In: American Sociological Review, 22 (6), p. 635.

Miles, B. T. / Greenwood, A. B. / Benito-Alifonso, D. / Tanner, H. / Galan, M. C. / Verkade, P. / Gersen, H. (2017): Direct Evidence of Lack of Colocalisation of Fluorescently Labelled Gold Labels Used in Correlative Light Electron Microscopy. In: Scientific reports, 7.

Misakian, A. L. / Bero, L. A. (1998): Publication bias and research on passive smoking: comparison of published and unpublished studies. In: JAMA, 280 (3), p. 250–253.

Mulkay, M. / Gilbert, G. N. (1981): Putting Philosophy to Work: Karl Popper's Influence on Scientific Practice. In: Philosophy of the Social Sciences, 11 (3), p. 389–407.

Nicolaisen, J. / Frandsen, T. F. (2019): Zero impact: a large-scale study of uncitedness. In: Scientometrics, 119 (2), p. 1227–1254.

Nye, M. J.: N-Rays (1980): An Episode in the History and Psychology of Science. In: Historical Studies in the Physical Sciences, 11 (1), p. 125–156.

Pfeffer, C. / Olsen, B. R. (2002): Editorial: Journal of negative results in biomedicine. In: Journal of negative results in biomedicine, 1, p. 2.

Popper, K. R. (1992 [1935]): The Logic of Scientific Discovery. London.

Sayao, L. F. / Sales, L. F. / Felipe, C. B. M. (2021): Invisible science: publication of negative research results. In: Transinformação, 33.

Schlich, T. (1993): Making mistakes in Science: Eduard Pflüger, his scientific and professional concept of Physiology, and his unsuccessful theory of diabetes (1903–1910). In: Studies in history and philosophy of science, 24 (3), p. 411–441.

Schneider, J. / Di Ye / Hill, A. M. / Whitehorn, A. S. (2020): Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. In: Scientometrics, 125 (3), p. 2877–2913.

Schuol, S. (2020): Scheitern in der Wissenschaft. In: Wissenschaftsreflexion, Eds. M. Jungert; A. Frewer, E. Mayr. Paderborn, p. 135–160.

Sigl, L. (2016): On the Tacit Governance of Research by Uncertainty. In: Science, Technology, & Human Values, 41 (3), p. 347–374.

Simon, B. (1999): Undead Science: Making Sense of Cold Fusion After the (Arti)fact. In: Social Studies of Science, 29 (1), p. 61–85.

Sovacool, B. (2005): Falsification and Demarcation in Astronomy and Cosmology. In: Bulletin of Science, Technology & Society, 25 (1), p. 53–62.

Stanley, T. D. (2005): Beyond Publication Bias. In: Journal of Economic Surveys, 19 (3), p. 309–345.

Star, S. L. / Gerson, E. M. (1987): The Management and Dynamics of Anomalies in Scientific Work. In: The Sociological Quarterly, 28 (2), p. 147–169.

Star, S. L. (1999): The Ethnography of Infrastructure. In: American Behavioral Scientist, 43 (3), p. 377–391.

Sterling, T. D. (1959): Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance – or Vice Versa. In: Journal of the American Statistical Association, 54 (285), p. 30–34.

Stern, J. M. / Simes, R. J. (1997): Publication bias: evidence of delayed publication in a cohort study of clinical research projects. In: BMJ, 315 (7109), p. 640–645.

Suñé, P. / Suñé, J. M. / Montoro, J. B. (2013): Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. In: PloS one, 8 (1).

Teixeira da Silva, J. A. / Bornemann-Cimenti, H. (2017): Why do some retracted papers continue to be cited? In: Scientometrics, 110 (1), p. 365–370.

Teixeira da Silva, J. A. (2015): Negative results: negative perceptions limit their potential for increasing reproducibility. In: Journal of negative results in biomedicine, 14.

Timmermans, S. (2011): The Joy of Science: Finding Success in a "Failed" Randomized Clinical Trial. In: Science, Technology, & Human Values, 36 (4), p. 549–572.

van Raan, A. F. J. (2004): Sleeping Beauties in science. In: Scientometrics, 59 (3), p. 467–472.

Vuong, Q.-H. (2020): The limitations of retraction notices and the heroic acts of authors who correct the scholarly record: An analysis of retractions of papers published from 1975 to 2019. In: Learned Publishing, 33 (2), p. 119–130.

Wang, J. / Veugelers, R. / Stephan, P. (2017): Bias against novelty in science: A cautionary tale for users of bibliometric indicators. In: Research Policy, 46 (8), p. 1416–1436.

Watson, J. D. / Crick, F. H. (1953): Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. In: Nature, 171 (4356), p. 737–738.

Zuckerman, H. / Merton, R. K. (1973 [1972]): Age, aging, and age structure in science. In: Aging and society: A sociology of age stratification, Eds. M. Riley; M. Johnson; A. Foner. New York, p. 292–356.