Reflexivity and fragility

Abstract
Reflexivity is, roughly, when studying or theorising about a target itself influences that target. Fragility is, roughly, when causal or other relations are hard to predict, holding only intermittently or fleetingly. Which is more important, methodologically? By going systematically through cases that do and do not feature each of them, I conclude that it is fragility that matters, not reflexivity. In this light, I interpret and extend the claims made about reflexivity in a recent paper by Jessica Laimann (2020). I finish by assessing the benefits and costs of focusing on reflexivity.

Keywords: reflexivity; fragility; master model; contextual; prediction; Laimann

1. Introduction

What influences whether we can predict, explain, and intervene successfully in the human sciences? I compare two potentially relevant phenomena. The first, the topic of this special issue, is *reflexivity* – which is, roughly, when theorising about or studying a target itself influences that target. (More on reflexivity's definition later.) The second is *fragility* – which is, roughly, when causal and other relations are hard to predict, holding only intermittently or fleetingly. (More on fragility's definition shortly.) I argue, in a nutshell, that it is fragility rather than reflexivity that matters. While reflexivity does sometimes have an impact on optimal methodology, usually fragility has a much bigger one.

In section 2, I introduce the notion of fragility and outline its methodological consequences. In section 3, I explain why, methodologically speaking, fragility matters more than reflexivity, by going systematically through cases that do and do not feature each of them. In section 4, I use this to build on recent work on reflexivity by Jessica Laimann. Finally, in section 5, I return to reflexivity itself, and consider how a focus on it both helps and hinders scientific progress.

2. Fragility and its consequences

Imagine two worlds. In one, there is underlying order. Causal relations are stable and long-lasting; mechanisms, structures and functional dependencies persist across many cases; laws are unchanging. How best to investigate such a world? By uncovering these cogs and wheels of nature, confident that they will work widely. Knowledge of them, accumulated over many generations, is the route to remarkable power. Technological artefacts can work reliably, by being engineered to combine and exploit cogs and wheels without disruption. To explain an event is a matter of identifying some configuration of these cogs and wheels, and perhaps thereby of seeing how the same cogs and wheels underlie many other, superficially disparate, events too. The science in such a world is one familiar from textbooks and popular image. Beneath the messy imperfection around us stands a Platonic order, and finding this order is science's mission.

Now imagine a second world, this time one in which laws and causal relations are fragile, winking in and out like bubbles in a boiling soup. In this world, things are different. Just because one thing causes another over there, that does not mean it will cause it over here. Working hard to discover underlying cogs and wheels is no longer an efficient use of our energies. When explaining things, we are forced knee-deep into idiosyncratic local detail; no eternal laws, rather each time a new look. Artefacts lose their power because the building blocks on which they rely are fragile. Progress is still possible in this fallen world, and remains vital for our fortunes, but it is piecemeal and patchwork.

Our world is an interlocking mixture of these two worlds. But many target relations are fragile and this needs to be reckoned with (Northcott forthcoming).

Speaking less roughly, define a relation to be *fragile* if, in the salient circumstances, it is not predictable when it holds. (Reciprocally, to be *stable* is to be non-fragile.) As we will see, defining fragility in this way enables us to track what is methodologically important. Fragility can arise because a relation holds only locally, or intermittently, or with inconsistent strength – and because these variations cannot easily be predicted.

Predictability has both a subjective and objective aspect, and therefore so does fragility. Predictability is relative to our knowledge: in a deterministic world, for example, nothing is hard to predict for Laplace's Demon, even while, of course, many things may remain hard to predict for us humans. But it is also true that some relations are harder to predict than others for reasons external to us. The difference between chaotic and non-chaotic systems is one example. Relations may also be fragile because they rarely operate in isolation, or because they require advanced levels of knowledge to identify, or because they are difficult to observe. Whatever the cause, in many cases it is not realistic to render something predictable just by trying to learn more. In those cases, in practice, we must take a relation's fragility to be a given.

Fragility is not the same as *complexity*, although complexity often leads to it. Fragility is a property of relations, complexity a property of systems. That said, complexity has several different definitions, and if it is defined in terms of unpredictability then any fragile relation will inevitably have arisen from some 'complex' system or other. But on most definitions of complexity, fragile relations can arise in non-complex systems too, as in the World War One truces example (Section 3). Conversely – on any definition of complexity – some relations even in complex systems are not fragile: summers are predictably warmer than winters, for instance, i.e., seasonality predictably causes variation in temperature, even though the weather system is a paradigm of complexity.

2.1 Master-Model strategy

There are further nuances, but for our purposes the above understanding of fragility is precise enough. Turn to our main focus, which is what fragility implies for methodology. It is useful to begin with a simplified, benchmark case of a *non*-fragile relation. Suppose we want to predict the motion of a newly discovered moon. To do so, we apply a Newtonian two-body model of gravity, inputting the moon and parent planet's masses, positions, and motions. Something like this procedure is a staple of actual space exploration. Why does it work? The answer is stability: the Newtonian model that has been successful elsewhere can be assumed to apply to a new case, because gravity itself can be assumed still to be operating in the same way. Each time, just re-apply the same Newtonian master model. Call this the *Master-Model* strategy.

(By a model 'applying', I mean a model correctly representing a force or cause in the target system. So, when a model applies, it explains (at least partially), and guides interventions. Models can 'apply' in other ways too, but I will not explore that here.)

A major advantage of Master-Model is that it is still effective even in the face of *noise* – understood here as significant effects from disturbing factors that are not captured by our model. For example, the moon's motion may be deflected by gravity from a second moon, by impact with a comet, or (at least for a small moon maybe) by human interference. If so, because of these disturbing factors, the Newtonian two-body master model would no longer predict accurately. Nevertheless, the model would still reliably identify *one* of the factors influencing the moon's motion, namely the gravitational interaction between moon and planet. In this sense, the model would still explain partially (Northcott 2013). To explain fully, or predict accurately, we would have to add in the effect of unmodeled disturbing factors. This strategy – of developing a master model and then in specific cases adding in disturbing factors as needed – was already advocated by Mill almost two centuries ago (1843). It has been a staple of philosophy of science about modelling, as many authors have focused on how models – even if idealized and even in the face of noise – may nevertheless succeed by isolating stable causal tendencies or arrangements (e.g., Cartwright 1989, Mäki 1992).

In this way, a master model provides some understanding even in the many cases where empirical accuracy is imperfect. Such an achievement, according to this view, is *superior* to mere empirical accuracy. Why? Because empirical accuracy in any particular case requires taking account of every local factor, no matter how sui generis or transient. But on this view, what is of greater interest to science, as a pursuit of systematic knowledge, is those factors that generalize – which is just what a master model captures.

Master-Model relies on stability. A master model can serve as a reliable base onto which case-specific disturbing factors may be added, only when the relations it describes are stable. (Mill himself was well aware of this: he had in mind economics, where he thought core psychological tendencies such as seeking to increase one's own wealth are indeed stable in the required way.) In easy cases, warrant to apply a master model comes from empirical success here and now: the Newtonian gravity model, for example, is given warrant by successfully predicting the motion of the moon. But often there is no empirical success here and now, because of noise. Then, warrant can come only indirectly, by importing empirical success from elsewhere. For example, even when noise means it predicts badly here and now, still we are justified in thinking the Newtonian model has correctly identified one gravitational force at work. Why? Because of the model's empirical success elsewhere. But such indirect warrant is justified only when there is stability. In this example, it is only because gravity operates in the same way across cases that the Newtonian model's warrant from success elsewhere stays good over here.

In sum, the crucial thing for Master-Model is stability, regardless of noise. *Without* noise, a master model is empirically accurate across many cases only when the relations it describes are stable. *With* noise, meanwhile, while a master model is no longer empirically accurate, now we may retreat to Mill's strategy, confident that a master model does at least capture some of the factors present, even if there are additional disturbing factors too.

2.2 Contextual strategy

But there is an alternative approach – for when we face fragility. To introduce this alternative, imagine now a different moon example. This time, the 'moon' in question is a toy moon on a string, being carried by a child around a toy planet. How might we predict the motion of *this* moon? The best candidate here for a master model is something psychological, perhaps that children continue to do actions that they are enjoying. It predicts that, if the child has carried the toy moon around the planet for two ‘orbits’ happily, they will continue for at least another two orbits. This prediction will be right sometimes. Other times, it will not: perhaps the child gets distracted, interrupted, or bored, or perhaps they are following instructions in an online science class (two orbits only), or perhaps they are playing a game with a friend (take it in turns to hold the moon). The underlying problem is fragility: the relation behind the prediction of continuity does not hold reliably. Using just a single model is no longer effective.

What alternative strategy works better? Many different models are available, some relatively formal, others we might think of more as loose hypotheses or rules of thumb. Some models cover a child’s behaviour in each of the deviant scenarios above: when distracted, bored, in a school class, with a friend, and so on. Others cover plenty of further scenarios: when a child is tired, when they are interacting with a sibling, when they are affected by poverty or divorce or moving to a new house, and so on. The key is which of these many models applies in any particular case. To discover that requires much case-specific work, looking for contextual clues and triggers: the character of this child, the nature of this household, is the child tired late in the day, is the child hungry, is the child – or friend or parent – generally frustrated after a prolonged lockdown, is the weather bright and warm or is it grey and miserable, and so on. In short, exactly the things a parent considers when trying to explain or predict a child’s behaviour. Instead of a single master model, we choose from many different models, case by case.

Label this new methodological strategy, *Contextual*. Contextual is not against wide-scope models as such; on the contrary, the larger the available toolbox of such models, the better. Rather, Contextual implies two things. First, a change in balance: relatively more scientific effort must be devoted to local empirical investigation. Because no single model may be assumed always to apply, we must be more sensitive to the details of each case, in order to select wisely from our toolbox. Second, a change in how models are developed. They should not be developed a priori or in the abstract, relying on real-world stability to ensure that if they capture a relation in one place, they will therefore capture it elsewhere too. Instead, models must constantly be empirically refined, in turn by constantly applying them to real-world cases. Such constant refinement is the best way to make models empirically productive, and to learn in what circumstances they are likely to apply (Ylikoski 2019).

When relations are fragile, even though relevant *models* may be putatively wide-scope, the *explanations* derived from those models are typically narrow-scope. In other words, these explanations cover only one or a few situations rather than, like the Newtonian gravitational model, many. Why? First, because relations at the heart of an explanation will, if fragile, often not hold widely. Second, because, as noted, empirical warrant cannot be imported from elsewhere, which means that empirical accuracy is always required here and now. This forces us to consider all local causes, no matter how sui generis. In turn, this usually requires going beyond just those factors captured by a wide-scope model and instead delving into local details, in the manner of a historian.

This localist picture dovetails with contemporary theories of causal explanation, such as Woodward’s (2003), which are framed not in terms of general laws but rather in terms of invariance relations that may be of very limited scope. It also dovetails with much recent philosophy of science, which emphasizes the need for local work to know when and how we may apply our models (Cartwright 2019). It dovetails too with work by Sandra Mitchell, who emphasizes that many generalizations in biology and social science are not ‘stable’, by which she means they do not apply universally (Mitchell 2000). (I thank an anonymous referee for alerting me to this.) Although Mitchell does not define stability in terms of unpredictability, she does draw a methodological conclusion similar to Contextual, namely that supplementary local work is required to track when and why such non-universal generalizations will apply.

To clarify the difference between Master-Model and Contextual: in both cases, use is made of models that may apply to many situations, and in both cases, contextual work is required to estimate parameter values. The difference lies elsewhere. In the moon example, the moon’s position and velocity vary continuously. The Newtonian gravity model tells us not just the details of that variation, but also when to expect it. Master-Model works well. There is no ‘surprise’ variation in gravity’s influence *that requires knowledge from beyond the model* to predict: the inverse-square law itself does not vary unpredictably, and neither is it difficult to predict when gravity will be present. But in fragile cases, we get just such surprise variation. So, we further need to investigate each time whether a model applies in the first place – whether its relations have changed their forms, and indeed whether its relations any longer hold at all. That is, we need Contextual.

Sometimes a model applies in many different places. Discovering a new mechanism can therefore be valuable because it enlarges our toolbox of available models. In this way, there is still scope for context-general scientific achievement. But if a relation captured by a model is fragile, then we may never just assume that the model applies; that still needs to be established anew each time. We are no longer in Newton’s world, so to speak.

With fragility, explanatory warrant requires empirical accuracy here and now, as noted. But empirical accuracy is now harder: the toy moon’s motion is harder to predict than is the real moon’s. It is more difficult to know which model applies, and noise is ubiquitous. But that, as it were, is nature’s fault, not ours. Still, this is not a counsel of despair: we can get a decent grip on the toy moon’s motion sometimes, some predictions are more accurate than others, and some explanations are fuller and better warranted. It is up to us to find them.

Summing up: what matters methodologically is fragility. When target relations are stable, it is best to investigate via a single master model (assuming an accurate one can be found), such as a Newtonian model of gravity. This strategy is effective even when empirical accuracy is disrupted by noise. But when target relations are fragile, matters change. A shift of emphasis is required – towards contextual, historian-like sifting. Local investigation is required each time to discover which of many candidate models might apply, and any model selected needs to be empirically accurate here and now.

3. Reflexivity versus fragility

Turn now to reflexivity. Several definitions have been offered of reflexivity, and of related (or, as sometimes used, synonymous) notions such as reactivity and performativity. Ian Hacking (1995) made famous the notion of unstable kinds. Reflexivity means that human kinds (i.e., kinds concerning humans) are potentially altered by feedback effects, with each alteration of a kind potentially inducing reactions in the target that in turn feed back into a further alteration of the kind, and so on indefinitely. But reflexivity has also been defined, more simply, as when theorising in itself impacts on the objects of study, with unstable kinds being merely one possible side-effect of that. Other definitions have been offered too. Below, I articulate reflexivity in terms of unstable kinds, but for our purposes nothing important turns on that.

Reflexivity is clearly distinct from fragility. All definitions of reflexivity have in common that there is some causal relation between theorising and the target of that theorising. But there is no reason that this causal relation must always be unpredictable, and thus no reason that it must always be fragile – and when it isn't fragile, we will have reflexivity without fragility. There are also uncontroversial cases, meanwhile, the other way round – of fragility without reflexivity. I go through examples below. As it were, with fragility the bubbles in the boiling soup come and go, whereas with reflexivity the scientist (perhaps unwittingly) is dipping their spoon into the soup and actively stirring it.

To assess the relative methodological significance of reflexivity and fragility, I will work through the different combinations of the two, using the following 2x2 table of examples. For each example, I report detailed case studies already carried out by others.

Table 1: Reflexivity versus fragility

	Reflexive kinds	**Non-Reflexive kinds**
Fragile relations	**Contextual** Schizophrenia Autism	**Contextual** Invasive species World War One truces
Stable relations	**Master-Model** Domestic dogs Gender roles	**Master-Model** Newtonian theory Electric toothbrush

Begin in the bottom-left corner, with reflexivity but not fragility. Start with domestic dogs (Khalidi 2010). The kind 'dog' is reflexive: as Khalidi explains, with repeated rounds of breeding over perhaps 15,000 years, this kind has changed dramatically, both morphologically and behaviourally. Traits including tameness, obedience, teachability, shepherding, hunting, and certain physical characteristics, have been selected for, so that what were originally wild wolves became domestic dogs and then later the many different breeds of domestic dog today. Human interaction with the kind changed it. There were many rounds of Hacking-style looping effects as new kinds themselves stimulated new human breeding behaviours, which in turn fed back to change the kinds once more. The key relation underpinning this history is that between breeding and the evolution of dog traits. This relation is stable: we can (for the most part) reliably predict the (rough) impacts of breeding interventions, and this reliability is exploited by dog breeders all the time. Further, we can reliably explain these impacts of breeding by appealing to general Darwinian theory – one of nature's cogs and wheels. Master-Model works well here for all of predicting, intervening, and explaining. This is despite reflexivity, and because of stability.

The example of gender roles (Laimann 2020) teaches the same lesson: optimal method tracks fragility (or its lack), not reflexivity. Briefly, the kinds 'masculine' and 'feminine' are reflexive, being strongly influenced by how people conceive of them. But many relations involving them are stable. (Indeed, Mallon (2016) argues that reflexivity has in this case been a stabilizing force, pushing the kinds back into line, so to speak, if they show signs of changing.) As a result, we can reliably predict the impacts of many interventions, and can reliably explain them, by appealing to wide-scope theories – in this case, sociological theories of gender roles. These theories describe relations such as that being perceived as male causes an individual to be subject to certain expectations regarding behaviour and appearance, in turn causing that individual to satisfy those expectations. Because these relations are stable, Master-Model is successful.

Turn next to the upper-left corner of Table 1, which features both reflexivity and fragility. Schizophrenia and autism are two of Hacking's own examples. Begin with schizophrenia. According to Hacking (1999, 112-14), because of reflexivity schizophrenia has changed its properties several times. At the start of the 20th century, when schizophrenia was first named, by Eugen Bleuler, its main symptom was flat affect. Auditory hallucinations (i.e., hearing voices), by contrast, were considered a minor issue, not specific to schizophrenia but rather observed in many other psychiatric conditions too. They were not to be worried about, and not something to hide from the doctor. The result was that hallucinations became increasingly widely reported by patients, and by the time a formal list of 12 symptoms of schizophrenia was compiled by Kurt Schneider 30 years later, the kind had changed, with hallucinations being designated the main symptom. But then, after the war, schizophrenia evolved from something viewed indifferently even favourably, to become instead a diagnosis that people wanted to avoid. As a result, patients became less willing to report hallucinations. This led the definition of the kind to be changed again, as hallucinations were gradually de-emphasized once more as a diagnostic criterion (although they are still listed as one of the main symptoms). Schizophrenia is, thus, according to this account, an unstable kind because of reflexivity effects.

How will the schizophrenia kind change next? It is hard to know. How attitudes to mental illness and to different symptoms of schizophrenia evolve, and how treatment of mental illness evolves, have been and will be determined by social and political relations that are difficult to predict. After the war, for example, something caused auditory hallucinations to be perceived as more shameful, but the operation of this causal relation, whatever it was, was not predicted. The relevant relation is fragile. If it were not fragile then, like with the breeding of domestic dogs, Master-Model could get us the answers we want. But there is no such master model available that explains the past or will predict the future of the schizophrenia kind reliably – no equivalent to Darwinian theory. As a result, in-depth local investigation is required instead, such as the history that Hacking recounts. Prediction, intervention and explanation are not straightforward in this case. Fragility explains why. It, not reflexivity, is the difference between the schizophrenia and domestic dogs examples.

Similar remarks apply to autism (Hacking 1995). First named in 1938, this kind has subsequently varied greatly in its definition, as well as in theories of what causes it, and in its degree of stigma. Reflexivity effects are an important part of the story, according to Hacking. The history of autism is the result of many social relations swirling in the background. The fragility of these relations is revealed by the need for detailed local investigation each time to discover which of them apply, and thereby both to explain the kind's history and to predict its future evolution.

Turn now to the two right-hand boxes. Being cases of non-reflexive kinds, they would not usually feature in discussions of reflexivity. But optimal methodology differs between them, and this difference is revealing, because only fragility is varying between the cases, not reflexivity.

In the upper-right corner, there is fragility but not reflexivity. Consider invasive species, as reported by Alkistis Elliott-Graves (2016, see also 2018, 2019). The relevant kinds here are things like tree species, soil nutrients, islands, and lakes. In the context of species invasions, none of these kinds is unstable or reflexive. But as Elliott-Graves recounts, despite knowing several mechanisms behind invasions and their harmful effects, we cannot predict reliably which of them will apply, and therefore cannot predict realiably the outcome or scope of invasion events. The relevant relations are fragile.

Consider one of Elliott-Graves's examples: plant-soil interaction. There exist both positive (certain fungi, and nitrogen fixers) and negative (pathogenic microbes) potential feedbacks to plants from the soil. Evolutionary interaction tends to favour the negative feedbacks, with the result that plants tend to become better off in a new area. Does this pattern enable us to predict the fate of plant invasions? Alas, no. How quickly plants accumulate pathogens is critical to an invasion's success, and this in turn varies with several further, local factors, such as the relative abundance of invaders and native plants, and the predation climate. As a result, prediction is difficult. Even when a combination of invader and soil microbes seems perfect for an invasion to succeed, often an invasion fails nonetheless. The relation between plant-soil set-up and invasion success is fragile. Plant-soil feedback interactions have been modelled extensively, and the relative abundance within a community of all-native plants has been predicted successfully. But when it comes to invasions of new communities, predictions are no longer reliable.

Many rules of thumb explain, or partially explain, invasions sometimes. Examples include: that islands, especially small ones, are more susceptible to invasions than are mainlands; that temperate climates are more susceptible than the tropics; and that within a taxon, smaller animals are more invasive than larger animals. Within plant taxa, the following traits correlate with successful invasions: small seed size; phenotypic plasticity; allelopathy, i.e., producing biochemicals that impact on the success of other organisms; adaptation to fire; and, at different times in different places, small and large size, flowering early and late, and both dormancy and non-dormancy. And the following traits of communities are all correlated with being easy to invade: when humans facilitate the invasion (perhaps inadvertently); when the community is disturbed; lack of biological inertia (i.e., the ecological balance can change relatively easily); particular plant-soil feedbacks, as just discussed; when the supply of resources fluctuates; and, depending on context, both high and low diversity. Similar lists can be compiled for marine ecosystems, insects, vertebrates, and so on.

But none of these many rules of thumb explains or predicts invasions reliably. They are like the various models in the toy-moon example: they are all fragile. For any given invasion, extensive case-specific work is required to work out which rules of thumb apply. Master-Model does not work.

So, epistemic difficulty occurs despite the lack of reflexivity. The invasive species case also illustrates how fragility is not restricted to human sciences. (More on the scope of fragility in Section 4.)

The World War One truces are another example of fragility without reflexivity. Briefly: truces broke out spontaneously in many parts of the Western Front, despite constant pressure against them from senior commanders. What explains this remarkable and moving phenomenon? According to (Northcott and Alexandrova 2015), the Master-Model strategy, represented in this case by the Prisoner's Dilemma game, is not successful, because the historical details turn out to contradict the Prisoner's Dilemma account in many ways. Indeed, the Prisoner's Dilemma actively directs attention away from the factors that were actually significant, and that bear on other instances of co-operation. The only way to successfully explain the truces is by contextual work, as exemplified by investigations by historians. The relevant relations are fragile. For example, when British and German soldiers were stationed at the same place on the front for a prolonged period, would that lead to spontaneous truces developing? Sometimes yes, sometimes no. To know which, each time further investigation is needed. And the kinds here (war, soldier, truce) are not reflexive, at least not in the context of this case.

In the lower-right corner, finally, there is neither fragility nor reactivity. Without fragility, Master-Model can again succeed: capturing the cogs and wheels of nature is again the efficient route to prediction, intervention, and explanation. A paradigm case is Newtonian theory. This describes relations that are non-fragile – indeed universal – and succeeds dramatically. It is a similar story with technological artefacts. These are deliberately engineered to exploit relations, such as that between connecting a battery and a light turning on, that, in a deliberately shielded environment, are stable.

Newtonian theory's target kinds are typically non-reflexive, as are those of technological artefacts. But as with invasive species and with World War One truces, when it comes to choosing between Master-Model and Contextual, it does not matter whether kinds are reflexive or not. What matters is whether relations are fragile.

4. Laimann and beyond

Jessica Laimann's penetrating (2020) discussion of reflexivity shares many of the above emphases. I turn now to how an analysis in terms of fragility complements and adds to her discussion.

According to Laimann, our concern with reflexivity is ultimately epistemic and methodological. Unstable kinds in themselves are not necessarily a problem. Rather, what matters epistemically are the processes and mechanisms behind that instability: how well do we understand those? When we do understand them well, as with domestic dogs and with gender roles, we are able to predict, intervene, and explain satisfactorily.

Laimann writes: "Only when we *understand the mechanisms* that support patterns of change and stability among the members of a kind are we in a position to *provide accurate explanations and make inductive inferences* across a variety of contexts." (2020, 1056, italics added) The notion of fragility adds a new underpinning to the italicised phrases. What matters is whether the background relations are fragile. If they are *not* fragile, then, by definition of fragility, we will 'understand' them well enough to 'provide accurate explanations and make inductive inferences'. Laimann also writes: "The problem with human interactive kinds is not merely that the classified objects change, but that they change in ways *unforeseen by our extant theoretical understanding of the world*." (2020, 1051, italics added) Fragile relations, by definition, lead to changes that are unpredictable without supplementary knowledge, in other words precisely to changes that are 'unforeseen by our extant theoretical understanding of the world'.

A focus on fragility also clarifies why, for many purposes, it is relations that matter, not kinds. This is particularly clear in the case of causation. To causally explain requires us to identify a causal *relation*. Successful interventions, at least according to many theories of causation, also require identifying causal *relations*. And predictability, and thus the ease of successful intervention, tracks the (lack of) fragility of *relations* – by definition of fragility. Master-Model, it is made clear, is an appropriate strategy only for when *relations* are not fragile. That is why it works with domestic dogs and Newtonian theory, but not with invasive species or schizophrenia.

As Laimann points out, the question of how quickly *kinds* change is a red herring. For example, many bacteria change their nature very fast, but they can still be analysed successfully by Darwinian theory. Gender roles, in contrast, do not change at all because (according to Mallon) of stabilizing social effects, but we will nevertheless predict, intervene, and explain wrongly if the underlying social relations are not understood. What matters is not kinds but relations.

A central element of Laimann's paper is her argument that human kinds are often hybrid in a particular way. They have a dual nature: they can be understood in terms of the properties that explicitly define the category (the 'base kind'), but also in terms of the social position an individual occupies or social role the individual plays in virtue of being recognised as a member of that category (the 'status kind'). Laimann gives the example of sex as a biological base kind, versus gender as a social status kind. In much everyday and scientific speech, 'man' and 'woman' are hybrid kinds, encompassing both of these aspects. Often, the base kind is stable while the status kind is fragile (although perhaps not in the case of gender, as mentioned earlier).

The fact that many human kinds are hybrid can lead to two errors, according to Laimann. The first error she calls *biased conceptualization*. This is when the status element in a kind is ignored, with the result that, surprised by it, our predictions and explanations go wrong. For example, if schizophrenia is treated purely in terms of a specific symptom profile or purely as a neurological condition, then we would miss (according to Hacking) how people diagnosed as schizophrenic are singled out for particular expectations, opportunities, and treatments, and how this in turn leads to a change in the behaviour of schizophrenics and thus, ultimately, to a change in the definition of the kind itself. As Laimann says, if we conceive of a hybrid kind "solely in terms of the base kind, without considering the associated status, causal pathways associated with the status disappear out of sight." (1060) The resulting gap in our knowledge renders relations around the hybrid kind fragile. Just knowing schizophrenia's current definition in terms of symptoms or neurological features will leave us unable to predict future changes in the kind reliably. Another example of biased conceptualization concerns the kind 'unemployed'. Here, the base kind is a dry economic definition of being without paid work when available for it, while the status kind is the social stigma of not having a job. According to Laimann, neglect of the latter aspect impedes our understanding of unemployed people's inferior health outcomes.

The second error, according to Laimann, is simply *not understanding social status effects*. Social mechanisms are many and complex, and their effects, or whether they are even operating at all, are often difficult to predict. In other words, social status effects are often the products of fragile relations. Not understanding social status effects is not a conceptual error, unlike biased conceptualization. Rather, it is just that even when we recognise the true nature of hybrid kinds, still the social science required to study them can be difficult. Laimann gives the example of the rise of the gay rights movement in the USA after the Stonewall riots in 1969, which greatly and rapidly changed the status kind component of ‘homosexual’. This event was the result of a perhaps unique constellation of complex social and political processes. It was hard to predict and remains hard to fully explain.

A key point for Laimann is that neither biased conceptualization nor failure to understand social status effects, directly concerns reflexivity. Reflexivity is not the key feature. Rather, each shortcoming is at root a deficit in our knowledge of causal relations surrounding the social status aspect of hybrid kinds. If we had had this knowledge then, reflexivity notwithstanding, we could have predicted and explained successfully. I agree with Laimann here. Thinking in terms of fragility allows us to pinpoint exactly what this deficit in our knowledge of causal relations is.

Laimann convincingly and usefully shows one route – hybrid kinds – by which human sciences fall prey to fragility. We may add to her analysis by noting that there exist other routes to fragility too, which have nothing to do with hybrid kinds (World War One truces). And fragility is not unique to human sciences (invasive species). Besides invasive species, other instances of fragility in natural sciences arguably include many cases from ecology generally (Sagoff 2016), and indeed from field biology generally (Dupré 2012). Fragility is also ubiquitous in many medical treatments, in data science (Pietsch 2016), and in complex systems generally. (Reflexivity too is not unique to human sciences; in addition to domestic dogs, there are arguably other cases from biology as well (Cooper 2004).)

5. Reflexivity revisited

Does a focus on reflexivity help or hinder? On the positive side, reflexivity can be a useful indicator of fragility. Laimann’s mechanisms of biased conceptualization and of not understanding social status effects are two ways this can happen. As an indicator, though, reflexivity is not infallible. Sometimes, as with domestic dogs and gender roles, reflexivity comes without fragility; and other times, as with invasive species and World War One truces, fragility comes without reflexivity.

Being aware of reflexivity also brings a second benefit. It can alert us to specific social mechanisms, knowledge of which helps us to *reduce* fragility. A familiar case illustrates: the self-fulfilling prophecies behind bank runs. Banks’ cash reserves typically cover only a fraction of their depositors’ credit, so if all depositors demand their money at the same time, the bank faces a liquidity crisis and can go bust. In normal times, this does not happen. But rumours or reports that a bank is in trouble can spur all depositors, made worried about the bank’s solvency, to try to withdraw their money at the same time. In this way, mere rumours of trouble can cause actual trouble – even if initially they are false. This is reflexivity in action. The analysis of a target – in this case, the rumours regarding the bank’s solvency – itself influences that target. The positive thing for science is that knowing this mechanism of the self-fulfilling prophecy allows us both to predict and to explain the run on the bank. Many historical bank runs have been explained in this way, at least in part, such as the collapse of the Dutch Tulip mania, the British South Sea bubble, many American banks in the Great Depression and, more recently, the collapses in 2007 of Northern Rock bank in the UK and IndyMac bank in the USA. And knowing the mechanism of the self-fulfilling prophecy does more. It also guides us towards interventions that can stymie this mechanism and thereby keep banks stable, such as granting regulators the power to prevent deposit withdrawals, or guaranteeing all deposits up to a certain value. Such measures are designed to allay depositors' fears of a run on the bank losing them their money, thereby preventing the run in the first place, and so preserving the bank's solvency. Preventing bank insolvency in this way means that relations such as that between depositing money in a bank and being confident of having access to that money at a later date, are rendered reliable rather than fragile. That is, we are using knowledge of reflexivity to ensure stability.

A similar story is true of many other rational-expectations economic models. Knowledge of reflexivity enables us to make stable some relations that previously were fragile.

Is reflexivity an effect of fragility, or a cause of it? It can be either. Some fragile relations are causally upstream of reflexivity, others causally downstream. For example, in a bank run, will authorities intervene effectively? Suppose that this is hard to predict. Then the following relation is fragile: that a bank being rumoured to be in trouble causes the authorities to intervene effectively. In turn, because of this fragility, depositors lack reassurance, and so rumours of trouble can become self-fulfilling prophecies. That is, here fragility causes reflexivity. But matters do not stop there. For the occurrence of bank runs then causes a new relation to become fragile, as noted above, namely the relation between depositing money and being confident of having access to that money at a later date. Having itself been caused by one case of fragility, reflexivity then causes a new case of fragility.

That is the positive side, so to speak, of a focus on reflexivity: it can be an indicator of fragility, either as a cause or effect of it, and it can be a source of insight into social mechanisms. Turn now to the negative side.

The main danger is simply misdirection: we should focus on fragility, because fragility matters more. Reflexivity is a distraction.

But there is, in addition, a second danger: a mistaken scepticism about social prediction. For in its more radical forms, an emphasis on reflexivity denies that systematic predictive success in human sciences is possible at all. This scepticism is a priori. Given free will, the argument runs, humans are always free to react to any prediction about themselves in such a way as to falsify it. Suppose, for example, that an unpopular candidate is predicted to win an election because of low voter turnout for their opponent. This very prediction may then inspire the previously apathetic supporters of the opponent to come out and vote, thereby preventing the unpopular candidate from winning – and so the prediction falsifies itself. Because prediction about humans is always vulnerable to reflexivity effects in this way, it is inevitably unreliable. Therefore, the argument concludes, in human sciences we cannot use prediction to test scientific theories in the same way as we can in natural sciences. We cannot use it as a basis for action either.

This scepticism has had distinguished proponents. They include Hayek, Popper, and MacIntyre; many interpretivists; and various intellectuals outside academia, such as George Soros, Michael Frayn, and Jonathan Miller.

Some responses to this scepticism, such as Mill's, have claimed that free will is compatible with determinism. But whether it is or not, a better response, I think, is to point out the obvious fact that social predictions often are successful. The interesting thing is when and why they are.

No doubt, social prediction is challenging. Prediction generally is challenging. But from a methodological point of view, reflexivity is merely one source of fragility. It no more makes third-person causal investigation impossible than does any other source. As Laimann emphasizes, what matters is not that reflexivity moves the goalposts, but rather whether we understand the mechanisms behind the moving of the goalposts. If we do, then we can still predict perfectly well, as we do with domestic dogs and gender roles. Reflexivity does not somehow magically negate this. A priori arguments that it does are falsified by ample experience.

But it is not just that scepticism about social prediction is misguided. It is also pernicious. Why? Because, in effect, it seeks to deny human sciences the possibility of empirical testing, which is the key to advancing knowledge. Here is one example. (There are many others.) In a UK government press conference in April 2020, Health Secretary Matt Hancock was asked whether total UK Covid-19 deaths could still be kept below 20,000. (The official figure had just reached 10,000.) He replied: “The future path of this pandemic in this country is determined by how people act. That’s why it’s so important that people follow the social-distancing guidelines. *Predictions are not possible, precisely because they depend on the behaviour of the British people*.” (BBC 2020, italics added) Here, Hancock explicitly endorses the a priori scepticism about social prediction. At one level, his statement can be read simply as a statement of epistemic humility, correctly noting that part of the causal chain determining case numbers would be the public’s behaviour. But the statement also makes it impossible to hold policy to account. It is deemed that we cannot fairly assess whether a prediction of a particular policy’s effect is rational and, thus, whether the policy is worthy of praise or blame. No prediction can be deemed better than any other. Responsibility for the outcome is conveniently evaded – and, in this case, put on the public instead.

Of course, the context here was that Hancock wanted (understandably) to maximize public following of restrictions, and so had reason to emphasize the importance of that rather than of any particular prediction. Perhaps telling the public that the outcome depended on its own behaviour was merely in the service of this urgent practical imperative, and Hancock himself did not really believe that social predictions cannot be fairly evaluated. But what Hancock himself did or did not believe is beside the point. When it gives cover to such evasions of responsibility, the a priori scepticism is pernicious, both epistemically and morally.

Denying rational social prediction is, in effect, philistine. It denies that much actual, successful scientific inquiry is even possible. From where does this philistinism arise? A full answer is beyond the scope of this paper, but here is a speculation. The root cause of the error is a mistaken focus on reflexivity rather than fragility and, in turn, this mistaken focus reflects an agenda drawn primarily not from philosophy of science or from science itself, but instead from wider philosophy. This agenda is ultimately external. It can be seen in many traditional handbooks, anthologies and introductions to philosophy of social science, or in, to take one example, how David Papineau remembers and presents the agenda of philosophy of social science in (Papineau 2008). Typical questions are: ‘What is intentional explanation?’ and ‘How can we causally explain human action?’, which reflect the agendas of philosophy of action and philosophy of mind; ‘Is there collective agency?’, which reflects the agendas of metaphysics, ethics, and philosophy of action; and ‘Do special sciences reduce to physics?’, which reflects the agendas of metaphysics and philosophy of mind. The same questions drive much of the attention given to reflexivity. But none of them is primarily motivated by knowing what methods make human sciences successful or unsuccessful, where this is measured in the currency of predictions, explanations, and interventions. To know that, focus instead on fragility.

Acknowledgements
I am grateful for useful feedback, first, from the audience at the Reactivity, Prediction and Intervention in the Human Sciences conference in Helsinki in August 2021, and second, from two anonymous referees for this journal.

References

BBC (2020). 'Coronavirus: 'Sombre day' as UK deaths hit 10,000', BBC News online, 12th April 2020, [Coronavirus: 'Sombre day' as UK deaths hit 10,000 - BBC News]

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. (Oxford: Oxford University Press.)

Cartwright, N. (2019). *Nature, the Artful Modeler*. (Chicago: Open Court.)

Cooper, R. (2004). 'Why Hacking is wrong about human kinds', *British Journal for the Philosophy of Science* 55, 73-85.

Dupré, J. (2012). *Processes of Life*. (Oxford University Press.)

Elliott-Graves, A. (2016). 'The problem of prediction in invasion biology', *Biology and Philosophy* 31, 373-393.

Elliott-Graves, A. (2018). 'Generality and causal interdependence in ecology', *Philosophy of Science* 85, 1102-1114.

Elliott-Graves, A. (2019). 'The future of predictive ecology', *Philosophical Topics* 47, 65-82.

Hacking, I. (1995). 'The looping effects of human kinds', in D. Sperber, D. Premack and A. J. Premack (eds), *Causal Cognition: A Multidisciplinary Debate*, 351-394. (New York: Clarendon Press.)

Hacking, I. (1999). *The Social Construction of What?* (Cambridge, MA: Harvard University Press.)

Khalidi, M. (2010). 'Interactive kinds', *British Journal for the Philosophy of Science* 61, 335-360.

Laimann, J. (2020). 'Capricious kinds', *British Journal for the Philosophy of Science* 71, 1043-1068.

Mäki, U. (1992). 'On the method of isolation in economics', *Poznan Studies in the Philosophy of the Sciences and the Humanities* 26, 19-54.

Mallon, R. (2016). *The Construction of Human Kinds*. (Oxford: Oxford University Press.)

Mill, J. S. (1843). *A System of Logic*. (London: Parker.)

Mitchell, S. (2000). 'Dimensions of scientific law', *Philosophy of Science* 67, 242-265.

Northcott, R. (2013). 'Degree of explanation', *Synthese* 190.15, 3087-3105.

Northcott, R. (forthcoming). *Science for a Fragile World*. (Oxford: Oxford University Press.)

Northcott, R., and A. Alexandrova (2015). 'Prisoner's Dilemma doesn't explain much', in *The Prisoner's Dilemma*, ed. Martin Peterson, 64-84. (Cambridge: Cambridge University Press.)

Papineau, D. (2008). 'Five philosophy of social science answers', in D. Rios and C. Schmidt-Petri (eds) *Philosophy of the Social Sciences: Five Questions* (Automatic Press), 99-110.

Pietsch, W. (2016). 'The causal nature of modeling with Big Data', *Philosophy and Technology* 29, 137-171.

Sagoff, M. (2016). 'Are there general causal forces in ecology?' *Synthese* 193, 3003-3024.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. (Oxford: Oxford University Press.)

Ylikoski, P. (2019). 'Mechanism-based theorizing and generalization from case studies', *Studies in History and Philosophy of Science* 78, 14-22.