Synthese (2024) 203:114
https://doi.org/10.1007/s11229-024-04530-1

ORIGINAL RESEARCH

Confidence in Covid-19 models

James Nguyen[1,2,3]

Received: 8 July 2022 / Accepted: 5 February 2024

Abstract
Epidemiological models of the transmission of SARS-CoV-2 played an important role in guiding the decisions of policy-makers during the pandemic. Such models provide output projections, in the form of time -series of infections, hospitalisations, and deaths, under various different parameter and scenario assumptions. In this paper I caution against handling these outputs uncritically: raw model-outputs should not be presented as direct projections in contexts where modelling results are required to support policy -decisions. I argue that model uncertainty should be handled and communicated transparently. Drawing on methods used by climate scientists in the fifth IPCC report I suggest that this can be done by: attaching confidence judgements to projections based on model results; being transparent about how multi-model ensembles are supposed to deal with such uncertainty; and using expert judgement to 'translate' model-outputs into projections about the actual world. In a slogan: tell me what you think (and why), not (just) what your models say. I then diffuse the worry that this approach infects model-based policy advice with some undesirably subjective elements, and explore how my discussion fares if one thinks the role of a scientific advisor is to prompt action, rather than communicate information.

Keywords Models · Simulations · Covid-19 · Climate · Confidence · Objectivity · Science and policy

1 Introduction

During early 2020 governments across the world were faced with policy decisions that would impact the lives of billions. Covid-19 cases were found in multiple countries,

✉ James Nguyen
james.nguyen@philosophy.su.se

[1] Department of Philosophy, Stockholm University, Stockholm, Sweden

[2] Institute of Philosophy, School of Advanced Study, University of London, London, UK

[3] Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, UK

Published online: 02 April 2024

and evidence from countries where the disease was detected early, notably China and Italy, suggested that the disease could spread relatively quickly through a population, and swiftly overwhelm healthcare systems. At the time, although they lacked vaccines or treatments, policy-makers had various non-pharmaceutical interventions (NPIs) at their disposal: they could require case isolation; they could enforce social distancing for some, or all, of the population; and they could close schools and universities, for example. One of the primary questions they faced was whether they should allow the disease to spread across the population, perhaps mitigated, i.e. slowed by some targeted NPIs, or whether they should attempt to suppress transmission entirely via combining multiple NPIs, targeted at the entire population, including, as it turned out in many countries, a nationwide lockdown.

Epidemiological models were a major, if not the primary, source of evidence that fed into this decision (at least in the United Kingdom). Of particular relevance was Report 9 from a team at Imperial College London (Ferguson et al., 2020; SPI-M-O, 2020d). That report summarised the results of an agent-based simulation model, originally developed to represent the spread of influenza, parameterized to what was known about Covid-19 at the time.[1] The report included times-series of projected deaths, hospitalisations, and intensive/critical cases, under different assumptions about the severity of the disease and the NPIs utilised by the government. It demonstrated that in many of these projections, demand for hospital beds far exceeded those available at the time. The report was part of the UK's Scientific Pandemic Influenza Group on Modelling, Operational (SPI-M-O) sub-group's output, and was considered by the Scientific Advisory Group for Emergencies (SAGE) on 16 March 2020 (SAGE, 2022). Lockdown measures were introduced across the UK roughly a week later (BBC, 2020b).

The model that generated the results included in Report 9 was not a perfect replica of the UK population. The model didn't include care-homes for example, where Covid-19 had a disproportionate impact, at least in the early stages of the pandemic (Burki, 2020). And whilst agents in the model were associated with households, schools, and places of work, they didn't overcrowd supermarkets, or attempt to keep appropriate distances on public transport. They didn't make decisions about wearing, or not wearing, face masks (Boulos et al., 2023). Agents in the model weren't classified according to socioeconomic bands, or races, and so the NPIs in the model reduced contacts uniformly across individuals (with the exception of one NPI: social distancing of the elderly): there were no essential workers in the model. And again, we now know the relevance of these factors (Razai et al., 2021).

Even given the information available at the time, it was at the very least plausible that the model diverged from the target with respect to some causally relevant factors, factors which impacted the spread of the disease. It would be surprising then, even from a March 2020 perspective, if the model-outputs exactly matched how the pandemic would evolve in the UK, even assuming the model accurately represented the details of the disease and the government's response to it. In short, there were many model

[1] For presentations of the model targeted at influenza see (Ferguson et al., 2005, 2006) particularly the supplemental material. For discussion about its application to Covid-19 see (Ferguson et al., 2020; SPI-M-O, 2020d) and (RC Centre for Global Infectious Disease Analysis, 2020) (the latter being, essentially, the model used to generate the results in the former).

uncertainties; ways in which it was plausible that the model would diverge from its target.

In this paper I argue that these uncertainties were not communicated appropriately.[2] Drawing on Report 9, and discussions of the report in the freely accessibly SAGE minutes, I emphasise that the raw model-outputs—the projected deaths, demands on beds, etc. associated with each projection scenario—were offered for policy-support without appropriate reflection on the model-target distinction. Drawing on the fifth IPCC report, I provide two suggestions for how this could be done in the future: (i) the IPCC uncertainty framework allows for a much richer communication of scientific uncertainty, which would have been appropriately used in the case of Covid-19 models (Mastrandrea et al., 2010); and (ii) expert judgement can be used to adjust model-outputs into projections concerning the actual world that take into account the uncertainties associated with the models from which they originate (Stocker et al., 2013; Thompson et al., 2016). Both of these techniques require eliciting what modellers believe; in the first case how well their models represent the world, in the second what they directly believe about the world itself. Hence my slogan: tell me what you think (and why), not (just) what your models say.

One might worry about this recommendation. Both (i) and (ii) require that experts 'subjectively' evaluate, or adjust, their, to some extent 'objectively' derived (in the model) evidence. As such, they introduce a subjective, or at least non-mechanical, element into the process of offering model results for policy consideration. Whether this should be avoided depends on how one thinks about subjective and objectivity in science. Drawing on Douglas (2004), I argue that there are various senses of 'objectivity' according to which introducing aspects external to the model when offering modelling results for policy advice is legitimate, and uncritically reporting bare model-outputs is not.

I proceed as follows. Section 2 introduces the epidemiological model of interest in this paper, CovidSim, with a focus on clarifying the structure of the uncertainties associated with, and the projections extracted from, it. Section 3 introduces the best practices from climate science I am urging be used by epidemiologists. Section 4 discusses how these could have been utilised in the epidemiological context. Section 5 addresses the worry that these practices introduce subjective elements into the process of providing information for policy decisions, and considers whether adopting such practices would have undermined government action. Section 6 concludes.

2 Covid-19

In this section I introduce the CovidSim model and situate it within the context scientists and policy-makers found themselves during March 2020.

[2] My aim is not to criticise the details of CovidSim (at least directly), nor to proclaim about whether the evidence it provided was sufficient to justify the policies that were drawn up on its basis (Winsberg et al., 2020, 2021; van Basshuysen & White, 2021a, 2021b) My focus is rather on how it was presented to policy-makers.

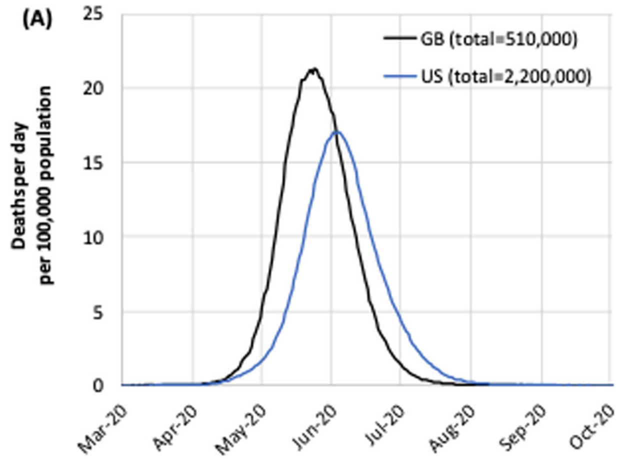

Fig. 1 Deaths per day per 100,000 in GB and the US, within the model, in an unmitigated scenario (Ferguson et al., 2020, p. 7, Fig. 1)

2.1 CovidSim and Report 9

The information presented in Report 9 was derived from an agent-based simulation model, originally developed to represent the spread of influenza, parameterized to what was known about Covid-19 at the time.[3] The model has a spatial structure. It consists of locations: households, places of work, schools, and so on. Agents in the model have a defined age, and are associated with households and other locations. They can be in one of three states: infected, susceptible, and recovered. Transmission events occur through contacts between susceptible and infected agents, dependant on their respective locations. Once an agent is infected they can transmit the disease to other susceptible agents (after an incubation period), and are either symptomatic or asymptomatic, with the former being more infectious than the latter. Depending on their age, a certain proportion of the infected agents are hospitalised, and a certain proportion of them require critical care, for a certain number of days. A certain number of infected agents (including those who do not require critical care) die from the disease.[4] By running the simulation for a period of time, the model produces time series of when and how many agents in the model die and require hospitalisation and critical care. Figure 1 demonstrates the projected number of deaths per day per 100,000 population in both Great Britain (excluding Northern Ireland) (GB) and the United States (US) within the model.

[3] For details of the model see the references in footnote 1. See (Maziarz & Zach, 2020) for a philosophical evaluation of agent-based epidemiological models in general.

[4] I am being deliberately vague about these ‘proportions’. Although not explicitly discussed in (Ferguson et al. 2020) it is stated that different severity scenarios (i.e. proportion of hospitalised cases requiring intensive care, and different infection fatality rates) were explored. In the version of the report considered by SAGE and SPI-M-O (which has minor differences to Report 9) uncertainty ranges for these figures are explicitly stated (SPI-M-O, 2020d).

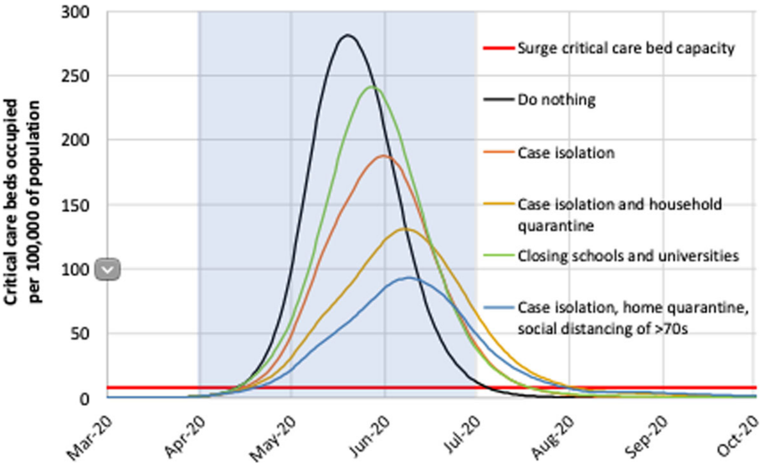

Fig. 2 Occupied critical care beds per 100,000 population in various scenarios in GB (Ferguson et al., 2020, p. 8, Fig. 2)

From this description of the model it should be clear that the time densities (and possibly total numbers) of deaths and hospitalisations depends on the frequency of the contacts between the infected and susceptible agents. Thus, one way to reduce these densities is to reduce the contacts between such agents. NPIs are methods for doing this. Five NPIs were modelled: Case isolation (CI) (according to which symptomatic cases stayed at home for a seven day period, reducing non-household contacts by 75% (but leaving household contacts unchanged), with 75% of households complying); Voluntary home quarantine (HQ) (all household members of a identified symptomatic case stay at home for 14 days, doubling household contact rate during that period but reducing community contact rate by 75%, with 50% of households complying); Social distancing of the over 70s (SDO) (reducing workplace contact rate by 50%, increasing household contact rate by 25%, and reducing other contacts by 75%, with 75% compliance); Social distancing of the entire population (SD) (all households reduce contact outside of household, school, and workplace by 75%, workplace contacts reduced by 25%, and household contacts increased by 25%); and the Closure of schools and universities (PC) (all schools closed, 75% of universities closed, household contact rates for student families increase by 50% and community contacts increase by 25%) (Ferguson et al., 2020, p. 6).

Each of these NPIs reduce peak incidences in the model (this is not to say that the peaks won't reappear once the interventions are lifted). Figure 2 depicts the impact of each NPI in isolation in place for three months from April through to July, in comparison to the 'do nothing' scenario over a three month period (note the line indicating the surge critical care bed capacity).[5]

[5] The Imperial team also considered the effect of turning the NPIs 'on' and 'off', based on some trigger condition during a simulation run. Trigger conditions that were considered included cumulative weekly ICU cases of 100, 300, 1000, and 3000 (Ferguson et al., 2020) Although not explicitly discussed in Report 9, it is stated that triggers that depend on per capita incidence rates were also explored.

Another important feature of the model is the relationship between the rate of infectious contacts (which depends on the infectiousness of the disease and the contact rate, assumed here to be what it would be in the absence of any NPI) and the mean infectious period of the disease. Based on these two values the (now infamous) basic reproduction number, R_0, estimates the expected number of secondary cases produced by a typical single infectious agent in a completely susceptible population. In CovidSim, $R_0 = 2.4$ was taken as a baseline assumption, although $R_0 \in \{2, 2.2, 2.4, 2.6\}$ were also explored.

From an analytic point of view we can distinguish between two different kinds of 'variables' within the model. First, there are *parameters* such as the percentage of infections requiring hospitalisation and/or critical care, the infection facility rate, and R_0. Second, there are *scenarios* such as whether, in which combination, and under which triggers, the NPIs are in operation in the model. Thus, multiple model runs are required to explore how the time -series data of deaths, hospitalisations, and peak ICU beds vary, depending on the values of these parameters and scenarios.[6]

So, the time -series outputs of the model depend on the way in which it is parameterised, and the NPI scenario under consideration. Figure 3 displays the how the model behaves under these differing assumptions.
The time -series presented include simulation runs based on uncertain parameters (e.g. different values of R_0 and different assumptions about the severity of the disease), and based on different scenarios (e.g. combinations of, and triggers for, the NPIs, henceforth I will include 'trigger' within 'scenario'). As a result, the different projections included in Report 9 reflect and communicate some level of parameter and scenario uncertainty. The authors also note that '[o]verall, we find that the relative effectiveness of different policies is insensitive to the choice of local trigger (absolute numbers of cases compared to per-capita incidence), R_0 (in the range 2.0–2.6), and varying IFR in the $0.25\% - 1.0\%$ range' (Ferguson et al., 2020, p. 8). So the report is fairly explicit (given what is communicated, and how it is described), that some sensitivity analysis has been performed, and that their recommendations do not change across the range of their uncertainties with respect to these parameters and scenarios.

But what's important for my current purpose is that for each parameter/scenario projection, the time -series extracted is the result of reporting the time -series data *within the model* (possibly averaged over multiple runs, see footnote 6). And as a result, there is no explicit discussion of how we should conceptualise the relationship between the time -series in the model, and the *actual* values of deaths, hospitalisations, and intensive care cases, we should expect to see, even assuming that the values of the parameters were correct, and the government implemented the appropriate scenario

[6] Third, it should also be noted that the model is 'stochastic' in the sense that it involves probabilistic factors; including the generation of the population and spatial structure, the initial infections within that population, and the variance in infectiousness between infected agents. As a result, a run of the simulation requires random seeds to initialise the model (although this process is governed by pre-specified probabilistic information). Different random seeds may induce different time -series data, even for the same parameterisation and scenario, so in general such simulations are ran multiple times. It is suggested that the model time series presented in Report 9 was the result of averaging over 10 different stochastic realisations (RC Centre for Global Infectious Disease Analysis, 2020). Whilst the stochastic nature of the model introduces important questions concerning the verification of the model (Eglen, 2020) I will put these aside for my current purposes.

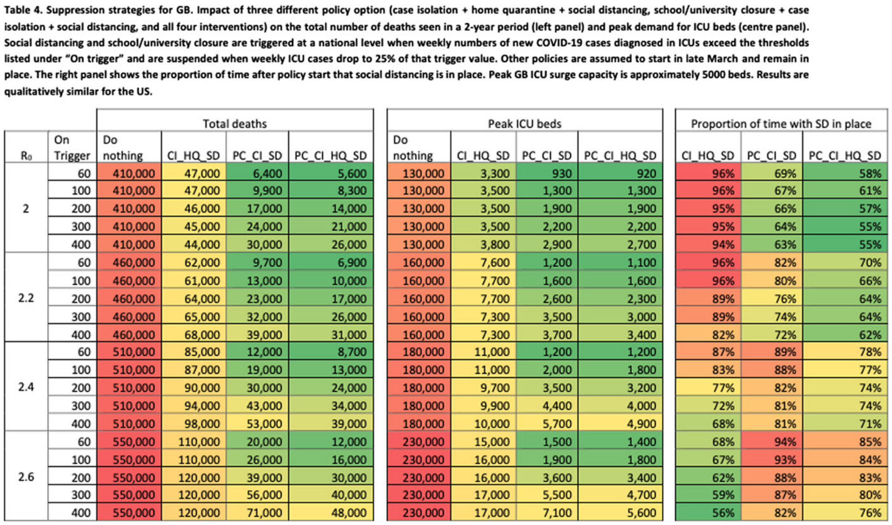

Table 4. Suppression strategies for GB. Impact of three different policy option (case isolation + home quarantine + social distancing, school/university closure + case isolation + social distancing, and all four interventions) on the total number of deaths seen in a 2-year period (left panel) and peak demand for ICU beds (centre panel). Social distancing and school/university closure are triggered at a national level when weekly numbers of new COVID-19 cases diagnosed in ICUs exceed the thresholds listed under "On trigger" and are suspended when weekly ICU cases drop to 25% of that trigger value. Other policies are assumed to start in late March and remain in place. The right panel shows the proportion of time after policy start that social distancing is in place. Peak GB ICU surge capacity is approximately 5000 beds. Results are qualitatively similar for the US.

		Total deaths				Peak ICU beds				Proportion of time with SD in place		
R_0	On Trigger	Do nothing	CI_HQ_SD	PC_CI_SD	PC_CI_HQ_SD	Do nothing	CI_HQ_SD	PC_CI_SD	PC_CI_HQ_SD	CI_HQ_SD	PC_CI_SD	PC_CI_HQ_SD
2	60	410,000	47,000	6,400	5,600	130,000	3,300	930	920	96%	69%	58%
	100	410,000	47,000	9,900	8,300	130,000	3,500	1,300	1,300	96%	67%	61%
	200	410,000	46,000	17,000	14,000	130,000	3,500	1,900	1,900	95%	66%	57%
	300	410,000	45,000	24,000	21,000	130,000	3,500	2,200	2,200	95%	64%	55%
	400	410,000	44,000	30,000	26,000	130,000	3,800	2,900	2,700	94%	63%	55%
2.2	60	460,000	62,000	9,700	6,900	160,000	7,600	1,200	1,100	96%	82%	70%
	100	460,000	61,000	13,000	10,000	160,000	7,700	1,600	1,600	96%	80%	66%
	200	460,000	64,000	23,000	17,000	160,000	7,700	2,600	2,300	89%	76%	64%
	300	460,000	65,000	32,000	26,000	160,000	7,300	3,500	3,000	89%	74%	64%
	400	460,000	68,000	39,000	31,000	160,000	7,300	3,700	3,400	82%	72%	62%
2.4	60	510,000	85,000	12,000	8,700	180,000	11,000	1,200	1,200	87%	89%	78%
	100	510,000	87,000	19,000	13,000	180,000	11,000	2,000	1,800	83%	88%	77%
	200	510,000	90,000	30,000	24,000	180,000	9,700	3,500	3,200	77%	82%	74%
	300	510,000	94,000	43,000	34,000	180,000	9,900	4,400	4,000	72%	81%	74%
	400	510,000	98,000	53,000	39,000	180,000	10,000	5,700	4,900	68%	81%	71%
2.6	60	550,000	110,000	20,000	12,000	230,000	15,000	1,500	1,400	68%	94%	85%
	100	550,000	110,000	26,000	16,000	230,000	16,000	1,900	1,800	67%	93%	84%
	200	550,000	120,000	39,000	30,000	230,000	16,000	3,600	3,400	62%	88%	83%
	300	550,000	120,000	56,000	40,000	230,000	17,000	5,500	4,700	59%	87%	80%
	400	550,000	120,000	71,000	48,000	230,000	17,000	7,100	5,600	56%	82%	76%

Fig. 3 Suppression strategies for GB (Ferguson et al., 2020, p. 13, Table 4)

(and the assumptions made about compliance by the public were accurate in each instance). In Report 9 at least, the time -series model-outputs were presented as if they referred directly to the (expected) actual values in the target (for example: '[i]n total, in an unmitigated epidemic, we would predict approximately 510,000 deaths in GB and 2.2 million in the US, not accounting for the potential negative effects of health systems being overwhelmed on mortality' (Ferguson et al., 2020, p. 7); '[t]able 3 shows the predicted relative impact on both deaths and ICU capacity of a range of single and combined NPIs interventions applied nationally in GB for a 3-month period based on triggers of between 100 and 3000 critical care cases' (Ferguson et al., 2020, p. 8)'; and '[o]ur results show that the alternative relatively short-term (3-month) mitigation policy option might reduce deaths seen in the epidemic by up to half, and peak healthcare demand by two-thirds' (Ferguson et al., 2020, p. 15).[7]

Another way of putting this, is that there is no reflection within the report on the possibility of *model-error*: plausible (even from a March 2020 perspective) ways in which the model may diverge from the actual target. And as stated above, there are various sources of potential model error that were plausible, even during March 2020. Hence my first recommendation:

Model Uncertainty Recognition: in contexts where there is reason to suspect that one's model may diverge, possibly severely, from one's target, one should, at the very least, recognise this by noting that the reported model-outputs should not be interpreted

[7] I discuss the inclusion of modifier phrases like 'approximately', 'might', and others like 'suggest', in Sect. 4.

as straightforward predictions/projections about the model's target. In slogan form: don't just tell me your models say.[8]

In a sense I take this recommendation to be obvious (although not always followed): it amounts to the straightforward advice to note that model-outputs are derived from the map, they do not immediately concern the territory. Moreover, we have already seen, it is commonplace for scientists to note the role of parameter (and scenario) uncertainty in determining their model results.[9] This recommendation simply asks them to go one step further: also note the role of model uncertainty.[10]

But whilst this is a useful starting point, simply adding such a disclaimer to a model result—'warning, this is a raw model-output'—doesn't go very far. Whilst policy-makers can be expected to understand the distinction between the model and the target, typically they are not well-positioned to understand the relationship between the two, and this is what is needed to get a grip on model uncertainty. Or to put it another way, since models aren't explicitly accompanied by 'keys' (or 'legends') in the way that maps are, simply providing policy-makers with the model (map) and noting that it is in fact a model (map), and not the target system (territory), is not particularly helpful. So whilst **Model Uncertainty Recognition** may be a necessary condition on communicating model uncertainty to policy-makers in a manner that is relevant to their decision-making, it does not seem sufficient (note here that I am currently assuming that policy- makers should be so informed, I return to this assumption in Sect. 5).[11] The next questions then, are first the extent to which **Model Uncertainty Recognition** was met in the context of Covid-19 more generally, and second, what more should we expect from modellers in their policy relevant communications? To answer these questions, it will be useful to move beyond Report 9 considered in isolation.

2.2 SAGE and SPI-M-O

Report 9 was written by a team of epidemiologists, with Prof. Neil Ferguson as the lead author, on behalf of the Imperial College London Covid-19 Response Team. At the time Ferguson was part of the SPI-M-O subgroup, and SAGE, which the subgroup

[8] And notice that caveats about parameter uncertainty, or emphasising that model-outputs are projections rather than predictions, since they survey multiple scenarios rather than assuming a particular one to be actualised, are not sufficient to do this.

[9] For more discussion on this, as it applies to CovidSim, see (Edeling et al. (2021) and Leung and Wu (2021) Notice that Edeling et al., 2021, p. 129 and supplementary Sect. 6) assert that: 'Model structure uncertainty is more fundamental' than the parameter uncertainty they investigate, but they do not explore it in their findings.

[10] There is however, a sense in which the latter kind of uncertainty stands apart from the former: whereas (provided sufficient computational time to run simulations for different values of the uncertainty parameters/different scenarios) parameter and scenario uncertainty can be investigated 'within the model', model uncertainty requires stepping outside the model and reflecting on its relationship to the target. But it is no less important for this reason.

[11] In fact, one could further argue that adding such a disclaimer *alone* and without emphasis could have the implicature that the appropriate key/legend to apply to the model is the identity key, according to which model results should be straightforwardly read as results about the target. Such an interpretation would amount to disingenuously assuming there is no model uncertainty whatsoever. I am grateful to an anonymous referee for encouraging me to be explicit about this.

fed into. The report was one of four papers upon which the 'consensus view' of SPI-M-O was based (SPI-M-O, 2020g). The other three papers were based on a non-age or spatially structured susceptible exposed, infectious, removed (SEIR) model (SPI-M-O, 2020e); a stochastic age structured model focusing on Buckinghamshire county (SPI-M-O, 2020c); and a non-age or spatially structured (but which differentiates between households) susceptible, exposed, infectious, detected, removed (SEIDR), with within and between household mixing, model (SPI-M-O, 2020b).

One way in which Report 9 already goes some way to meeting **Model Uncertainty Recognition** in a broader context, despite it not featuring in the report itself, can be found on the `.gov.uk` website where it is hosted. There it is stated that:

> '[t]hese results should not be interpreted as a forecast, but rather illustrative outputs under a set of assumptions to inform wider discussion. These modelling outputs are subject to uncertainty given the evidence available at the time, and dependent on the assumptions made'.

Important to note here is the instruction that the results not be interpreted as a forecast (despite the repeated use of the term 'prediction' within the report); the explicit recognition of the uncertainty facing the results; and the fact that the outputs are intended to 'inform wider discussion'. If we assume that this statement accompanied any presentation of the results from Report 9 (which is a generous assumption to make, note that it is not explicit in the slides accompanying the report (SPI-M-O, 2020a), which doesn't even recognise that the results presented were derived from a simulation model, let alone provide any detail about the model or model uncertainty), then this provides some measure of meeting **Model Uncertainty Recommendation**.

However, this is unsatisfactory for two reasons. First, it fails to differentiate between model uncertainty and parameter/scenario uncertainty. Given that the latter two are explicitly discussed in the report, it's plausible to interpret the accompanying warning as referring to this form of uncertainty only, with the implication that *if* the Imperial team were 'right' about these (i.e. the actual values for the disease parameters were within the ranges explored, and the government were able to implement some combination of NPIs in the manner assumed in the model), *then* the model-outputs corresponding to those model runs could be straightforwardly interpreted as predictions for the target system. Such an interpretation would ignore model uncertainty.

Second, and as preempted at the end of the previous section, even if the warning refers to model uncertainty (i.e. that the 'set of assumptions' referred to includes assumptions contained in the details of the model's structure and/or dynamics), the warning itself doesn't do anything other than assert the existence of such uncertainty. Without information about how this uncertainty impacts the possible relationship between the model-outputs and the values that we might expect to see in the target, decision-makers have nothing to go on. Recognising model uncertainty is one thing, providing information about its possible impact, and how to deal with it, is another. I return to this in Sects. 4 and 5.

There is another aspect of the warning, and the context in which Report 9 is embedded, that is also worth discussing. Recall that the stated purpose of the report was to 'inform wider discussion'. As noted, the report was one of four papers that fed into the SPI-M-O's 'consensus view', which was then the main input into the SAGE meeting on 16 March 2020. And the other three papers were all, in different ways, based on models that differed structurally from CovidSim (particularly, none of them were resolved at such a fine grained level of detail concerning the age and spatial structure of the agents within the model). As a result, one could interpret the strategy of focusing on *multiple* models, based on various different assumptions about the structure and dynamics of the target, as a recognition of the model uncertainty accompanying any individual model. Where these models agree may be thought to be independent of any particular detail of any particular model, and thus model -uncertainty is 'washed out' by considering multiple models. Thus, one can interpret the move from the multiple different models to SPI-M-O's consensus view as taking the form of robustness reasoning (Weisberg, 2006; Kuorikoski et al., 2010; Parker, 2011; Schupbach, 2018).

The consensus view found in (SPI-M-O, 2020g) supports this interpretation. It consists of five claims. First: a combination of case isolation, home isolation, and social distancing of vulnerable groups is deemed 'very unlikely' to prevent critical care facilities from being overwhelmed. Second: it is 'unclear' whether the addition of general social distancing on top of the aforementioned NPIs would reduce the reproduction number to less than 1. Third: adding general social distancing and school closures to the NPIs mentioned in the first point will 'likely' control the the epidemic if kept in place for a long period, and it is advised that this strategy should be followed as soon as practical (at least in the first instance).[12] Fourth: alternating periods of more and less strict social distancing measuring will 'plausibly' be effective at keeping critical care cases within capacity. Fifth: triggers could be enacted and lifted at levels of UK nations and regions, with duration of periods being less important than the number of contact reductions (and that there would be a two-three week lag between between measures being put into place and their impact being felt by ICUs).

For my current purposes, it is crucial to note that the language used in (SPI-M-O, 2020g) (in contrast to Report 9) is qualitative rather than quantitative: it is deemed 'very unlikely', 'likely', or 'plausible' that NPIs will have certain effects, the latter of which are not explicitly quantified.[13] I take it that this imprecision supports the idea that the consensus view should be interpreted as the results of robustness reasoning: moving from the quantitative outputs offered by each of the individual models to an imprecise qualitative description that captures (the decision relevant aspects of) where they agree. This implicitly recognises the potential impact of model uncertainty associated with any individual model, and goes some way to do so in a manner that can usefully inform policy-makers. On this (somewhat generous) reading then, (SPI-M-O, 2020g) does more than just meet the **Model Uncertainty Recognition** condition, it explicitly attempts to accommodate such uncertainty, via the introduction of multiple models,

[12] It is striking that this consensus statement includes an explicit normative recommendation for which NPIs *should* be implemented. For more on the role of scientists in making normative recommendations see (Rudner, 1953; Steele, 2012). Particularly relevant in this context is (Birch, 2021), to which I return in Sect. 5.

[13] Although there is no explicit recommendation for how the qualitative language should be interpreted and understood, which is particularly concerning given the well-known ambiguities in how people interpret these qualitative descriptions of probabilities (Willems et al., 2020) The question of how such interpretive guidance could be offered is returned to in the following section.

robustness reasoning, and the resulting shift to qualitative language can be thought of as the result of 'translating' raw model-outputs to claims that can be reasonably applied to the target system itself. So we can say that (SPI-M-O, 2020g) meets the following recommendation:

Model Uncertainty Accommodation: in contexts where there is reason to suspect that one's model may diverge, possibly severely, from one's target, one should reflect on, and communicate, how this may impact the relationship between the model-outputs and the values one expects to see in the model's target (possibly by discussing how the model-outputs compare to the outputs of other models within an ensemble). In slogan form: tell me what you think, not (just) what your models say.

Hopefully it is clear that accommodation is richer than recognition—the latter flags the model (map) vs. target (territory) distinction, the former provides policy-makers with a way to deal with it. However, as stated at least, it provides little constraint on the details of *how* scientists should perform such accommodations in the context of communicating with policy-makers. This concern can be applied to (SPI-M-O, 2020g): there is no explicit discussion of *how* the views reported there were arrived at. There is no mention of the details of any model that fed into the consensus (or indeed that they were based on modelling endeavours in the first place), or the method from which the multiple model results were taken to support the resulting consensus view. Both the original sources of evidence for the consensual position, and the methods at arriving at them, are thus 'black-boxed'.

Taking this together then, it is plausible that considering Report 9 in context, particularly with its accompanying warning, and understanding it as an ingredient in, rather than the sole contributor to, SPI-M-O's consensus statement goes some way to alleviating the worry that model uncertainty was ignored. However, some concerns remain, even under this understanding of the historical situation. Particularly, there is a lack of *transparency* regarding (i) the potential impact of model uncertainty on the individual model-outputs, i.e. the model data presented in the previous subsection of this paper, and (ii) the method from which the consensus statement is arrived at, given the inputs, i.e. the four modelling papers discussed above. Regarding (i): even if this data isn't supposed to be interpreted as providing predictions about the actual world (in the parameter and scenario regimes under consideration), it is clearly supposed to provide some guidance in this regard. Explicitly recognising this, even within Report 9 itself, would thus seem valuable. This is especially the case given the plausible division of labour involved (Prof. Neil Ferguson's simultaneous authorship of Report 9 and membership of SPI-M-O and SAGE notwithstanding): presumably the individuals who are best placed to reflect on how their results may be impacted by model uncertainty are precisely those who are most familiar with the details of the model itself (I return to this assumption in Sect. 4.2). Regarding (ii): whilst it is plausible that some form of robustness reasoning can be a successful way of accounting for (individual) model uncertainty (just as sensitivity analyses can account for parameter uncertainty), the details of how this is carried out—the method employed, the relative confidence attached to the individual models involved in the process, and so on—may have had a significant impact on which results were reported as the consensus view. Without knowing the details of how the view was developed, and how the individual models (with their associated model uncertainties) fed into this process, we are unable to assess whether model uncertainty was handled appropriately, even if, as a matter of fact, it was. The first concern compounds the second: reflecting on how model uncertainty could impact the results of CovidSim itself would provide a richer source of information to then input into the process of developing the consensus position. This leads to my third recommendation:

Transparent Model Uncertainty: Accommodation in contexts where there is reason to suspect that one's model may diverge, possibly severely, from one's target, one should reflect on, and communicate, how this may impact the relationship between the model-outputs and the values one expects to see in the model's target (possibly by discussing how the model-outputs compare to the outputs of other models within an ensemble), and this should be done in a transparent way. In slogan form: tell me what you think (and why), not (just) what your models say.

2.3 The model(s), or the communication?

Before investigating this recommendation in more detail, it's worth briefly clarifying the conceptual structure of my discussion in this section so far. I began by talking about the model uncertainties in CovidSim itself, with a focus on how they were handled in Report 9, before shifting to discussing how the epidemiologists involved did, or should, communicate with policy-makers, as exemplified by SPI-M-O's consensus statement. One could reasonably ask here, whether my recommendations are directed at the Covid-19 *modelling*, or Covid-19 *science communication*, particularly to policy-makers.[14] My answer, which I think helps distinguish the current contribution from much existing literature, is both.[15]

To make this point it is useful to draw on the literature on scientific representation, specially accounts of how models represent their targets (see Frigg and Nguyen, 2020 and Nguyen and Frigg, 2022 for a overviews); particularly, accounts associated with Giere (1988; 2004; 2010) and Frigg and Nguyen (2018). According to the former in order for a model to represent its target, it needs to be accompanied by *theoretical hypothesis* that specifies in which respects, and to which degree, the model is proposed to be similar to its target. If the hypothesis is true, and the model and the target are so similar, then the model is accurate in those respects and to that degree. According to latter, models, like maps, are, implicitly or explicitly, accompanied by *keys* (or *legends*) that associate aspects of models with aspects that their targets are proposed to have. If the target has such proposed feature, then the model is accurate with respect to it.

Whilst these are competing accounts, they agree that *without* a theoretical hypothesis or a key, a raw model-output, which is a report of the behaviour of a model, is just that: a fact about a model. It tells us nothing about the model's target system, e.g. projections about Covid-19 in Great Britain and the United States of America. So without a theoretical hypothesis, or key, reported model-outputs, like those found in Fig. 3 refer

[14] I'm grateful to an anonymous referee for encouraging me to be explicit about this.

[15] I have in mind here (Keohane et al., 2014; Schroeder, 2022) who focus primarily on the *communication*. For similar approaches that tie to this to modelling proper, see e.g. (Winsberg 2012; Parker 2014).

only to the behaviour of the model in question, in this case the agents in CovidSim.[16] Without such accompanying discussion, raw model-outputs, conceptually at least, tell us *nothing* about their target systems.

This holds independently of whether the modelling endeavour is being used in the context of communicating to policy-makers (and thus my arguments in favour of the above recommendations apply to the modelling context proper, not just the science communication context). For the most part, however, in modelling contexts, one can assume that the theoretical hypothesis/key that accompanies a model is, in some sense 'implicit' in the surrounding practice: modellers are generally sensitive to the map-territory distinction, have some awareness about the limits and scope of their models, and this impacts how they are interpreted (one might say that learning this is implicit in training to be a modeller). And so in such contexts a lack of explicit reflection on how model uncertainty may impact the relationship between the model-outputs and the target behaviour is not particularly egregious. But this is not the case in the context of communicating to policy-makers. Since they cannot be expected to be familiar with the interpretive conventions associated with e.g. epidemiological models there is increased pressure to make the theoretical hypothesis/key explicit in those contexts. And since the ultimate goal in those contexts is to communicate something like the all-things-considered epistemic information relevant to making an informed decision' (cf. Schroeder, 2022) (I return to this assumption in Sect. 5.2), the theoretical/ hypotheses/ key in question should take into account the known and suspected model-target divergences, i.e. model uncertainty, since this uncertainty is, I take it, clearly relevant in this regard. This motivates the recommendations offered in this section.

3 Lessons from climate science

Back to my recommendations. Something like **Transparent Model Uncertainty Accommodation** is at the heart of the IPCC's framework for handling and communicating uncertainty. The IPCC presents periodic 'assessment reports', designed to summarise the current state of knowledge about climate change and provide this much needed information to policy-makers. Whilst there is widespread agreement that anthropogenic climate change is real, when it comes to more fine-grained questions, residual uncertainty remains. As such, the IPCC requires some framework for scientists involved in writing the reports to conceptualise and report their uncertainty, in a way that allows for consistency across the author teams, and best reflects the underlying uncertainty itself. The fifth assessment report (AR5) was thus accompanied by a 'Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties' (Mastrandrea et al., 2010).

In this section I explore two aspects of this note and how they were utilised in AR5. The first concerns the way they communicate uncertainty on two metrics: *confidence*

[16] One could provide a hypothesis or key according to which the target has the exact same features that the model has, and thus justify exporting model-outputs to the world directly (and perhaps this sort of key is implicit in Report 9, if read in isolation) see (Frigg and Nguyen (2020, Chapter 4), for a critical discussion. This would amount to representing the target as if there were no model uncertainty whatsoever.

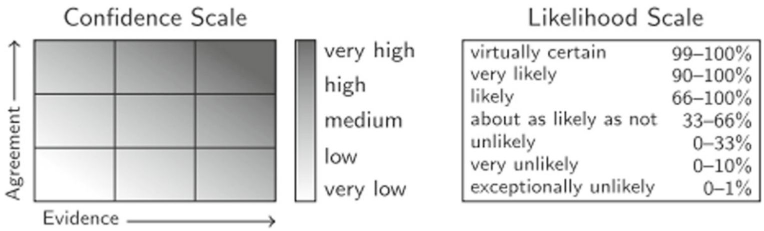

Fig. 4 IPCC uncertainty metrics, (Helgeson et al., 2018, p. 518), adapted from (Mastrandrea et al., 2010, p. 5)

'in the validity of a finding, based on the type, amount, quality, and consistency of evidence' (Mastrandrea et al., 2010, p. 1), and *likelihood*, i.e. '[q]uantified measures of uncertainty in a finding expressed probabilistically' (Mastrandrea et al., 2010, p. 1). The second concerns the fact that 'expert judgement' is included alongside models (and theory, mechanistic understanding, and so on) as a source of evidence, coupled with the observation that this judgement is used to transform model-outputs into target projections.[17] I demonstrate how each of these techniques could be employed in the epidemiological context.

3.1 Communicating uncertainty

The first metric used by the IPCC is 'confidence'. It is a qualitative metric, ranging from *very low* to *very high*, based on assessments of the underlying *evidence* (its type, amount, quality, and consistency), and the level of *agreement* between multiple sources of evidence. The second metric used by the IPCC is 'likelihood', which corresponds to our usual understanding of probabilistic statements. Recognising the fact that the relevant probabilities are imprecise, qualitative likelihood terms are attached to certain ranges of probabilities corresponding to some outcome. For example, if it is judged 66–100% probable that an outcome will occur, then this outcome is said to be *likely*. If it is 0–10% probable that an outcome will occur, then this outcome is said to be *very unlikely*. And so on. Figure 4 summarises these metrics.

Various statements in the report involve combinations of both notions, e.g.: 'Equilibrium climate sensitivity (ECS) is *likely* in the range 1.5 °C to 4.5 °C (*high confidence*)' (Stocker et al., 2013, p. 16, original emphasis). According to calibration of the likelihood and confidence scales, this is to be interpreted as saying there is a 66–100% probability (i.e. it is *likely*) that ECS is in the $[1.5 - 4.5\,^{\circ}\mathrm{C}]$ range, and that the evidence and/or agreement for this claim is *high*. Indeed every likelihood statement found in the report should be understood as being implicitly qualified as held to high or very high confidence, given that the guidelines recommend offering likelihoods only where such confidence levels are met (Mastrandrea et al., 2010, p. 4).[18] So many statements about

[17] For philosophical discussions of these aspects, and the relationships between them, see (Bradley et al. 2017; Helgeson et al. 2018).

[18] This recommendation isn't consistently followed in AR5, where probabilities are occasionally reported accompanied by e.g. a *medium* confidence qualifier (see, e.g. Stocker et al. 2013, p. 20).

the earth’s climate in the IPCC report are associated with two forms of uncertainty: they contain probabilistic content, and they are qualified by confidence judgements.

The question then, is what criteria are being drawn upon to provide information? Let’s start with the confidence metric. In broad terms, confidence is measured on two sub-dimensions: *assessment* of the evidence invoked in support of a claim, and the level of *agreement* between multiple sources of evidence for that claim. The sources of evidence include ‘mechanistic understanding, theory, data, models, [and] *expert judgment*’ (Mastrandrea et al., 2010, p. 1, emphasis added), and authors are told that they should:

> ‘Be prepared to make *expert judgments in developing key findings*, and to explain those judgments by providing a *traceable account*: a description in the chapter text of your evaluation of the type, amount, quality, and consistency of evidence and the degree of agreement, which together form the basis for a given key finding. Such a description may include standards of evidence applied, approaches to combining or reconciling multiple lines of evidence, conditional assumptions, and explanation of critical factors. When appropriate, consider using *formal elicitation methods to organize and quantify these judgments*’ (Mastrandrea et al., 2010, p. 2, emphasis added).

Notice that ‘expert judgement’ plays a dual role the process. One role for expert judgement is evidential, alongside evidential sources like theory or models (I address this role in the next subsection). The second is that expert judgement is drawn upon to assess the quality and/or agreement between these other lines of evidence. Claims in the report are based on multiple sources of evidence (models, theory, mechanistic understanding, etc.), and the authors of the report are encouraged to provide their assessment of those sources of evidence, and the level of agreement between them, in order to provide confidence qualifiers for those claims. These assessments are *subjective*, in the sense that they correspond to individual assessments of these factors, as evidenced by the recommendation that formal elicitation methods are suggested as tools for organising and quantifying them (although they are to ultimately combined into an author-team assessment). Author-teams are then encouraged to provide a ‘traceable account’ of how they arrived at these judgements, and how they reached agreement about the reported confidence level.

3.2 Model-to-world inferences

Let’s put confidence aside for the moment, and focus on how expert judgement can play a direct evidential role. As mentioned previously, ‘expert judgement’ is included as a legitimate source of evidence within the uncertainty guidelines. In cases where probabilities can be given, the guidelines say that ‘[l]ikelihood may be based on statistical or modeling analyses, *elicitation of expert views*, or other quantitative analyses’ (Mastrandrea et al., 2010, p. 3, emphasis added), and this is reiterated throughout the guidelines. e.g.: where a ‘range can be given for a variable, based on quantitative analysis or *expert judgment*: Assign likelihood or probability for that range when possible' (Mastrandrea et al., 2010, p. 4, emphasis added). So expert judgement is clearly recognised as a potential source of evidence for likelihood claims.

To see this in action, consider a claim discussed in detail by Thompson et al. (2016). Expert judgement is used in AR5 to *adjust* projections derived from formal modelling analyses. The 'Summary for Policy-Makers' includes table SPM.2 that reports projections for global mean surface air temperature and sea level changes under different emission scenarios (relative to the reference period 1986–2005).[19] The projected temperature change, for projection scenario RCP8.5, is reported as *likely* (i.e. $>$ 66%) to be in the 2.6–4.8 range by 2081–2100 (at least *high* confidence).[20] How do the authors arrive at this claim? They adopt a two-fold methodology. They first consider the behaviour of models in the Coupled Model Intercomparison Project Phase 5 (CMIP5) model ensemble. They use these models to calculate a 5–95% range for model temperature change (corresponding to the idea that there is a 90% probability, i.e. it will be *very likely*, that a model run will be in that range, for the RCP8.5 emissions scenario). Then: '[t]hese ranges are then assessed to be *likely* ranges after accounting for additional uncertainties or different levels of confidence in models' (Stocker et al., 2013, p. 23, SPM.2). As Thompson et al. (2016) note, this second step is the result of recognising that state-of-the-art climate models share systematic biases, and include numerous idealisation assumptions required for them to be tractable. As a result, an interval which is assigned >90% probability in 'model-land', is assigned only >66% probability in the actual world. The model-derived probabilities are downgraded to reflect expert judgement about the ways in which the models diverge from their targets.[21] The result of this process can be conceptualised as exactly the sort of thing I discussed in Sect. 2.3: expert judgement is used in establishing and making explicit the theoretical hypothesis or key that accompanies the models in question. Such judgement specifies the level of grain at which the model and target should be taken to resemble one another, or, putting it another way, it specifies how model-outputs should be translated to target projections.

4 From climate to Covid-19

In this section I first outline how the ways in which the IPCC handle and communicate uncertainty could have been utilised in the context of Covid-19, before offering some defence of the IPCC-to-Covid strategy I employ.

[19] These scenarios are called Representative Concentration Pathways (RCP). In AR5 four, RCP2.6, RCP4.5, RCP6, and RCP 8.5. are considered as possible greenhouse gas concentration trajectories.

[20] The confidence assessment is not made explicitly, but it is noted that the associated projections for the 2046–2065 period are medium, and that this is lower than for the 2081–2100 period, 'because the relative importance of natural internal variability, and uncertainty in non-greenhouse gas forcing and response, are larger than for 2081–2100' (Stocker et al., 2013, p. 23, SPM.2).

[21] It's worth highlighting that this is not an artefact of the example chosen; it's not that global temperature and sea levels are particularly troubling values to project. Their global nature means we have more confidence in projections for them, than we do more local variables.

4.1 Learning lessons

How could the IPCC's methods have been utilised in the epidemiological context? Recall that **Model Uncertainty Accommodation** recommends taking into account how model uncertainty may impact the relationship between model-outputs (considered in isolation or in combination) and, given this, what we expect, all things considered, to happen in the models' target, and **Transparent Model Uncertainty Accommodation** that this be done in a transparent manner. The practices of the IPCC suggest ways in which Report 9 and the consensus statement could, and should, have met these recommendations.

Taking Report 9 first. If the IPCC's framework is correct, then in addition to providing the time -series outputs derived from CovidSim, the report should contain a discussion of the the level of *confidence* the authors attach to their findings. These confidence judgements should measure the (perceived) quality of the evidence in favour of the claims, i.e. the perceived quality of how accurately CovidSim represents its target system(s). Such confidence assignments would reflect the idealisation and tractability assumptions contained in the model, and the model -uncertainty that accompanies them. And since the authors are in the best position (more on this in the next subsection) to evaluate the impact that these assumptions had on the relationship between the model-outputs and the values we should expect to see in the target (even assuming that the parameterization and scenarios in question were represented accurately) the confidence that they attach to their model would provide useful information to SPI-M-O and SAGE, policy-makers, and the wider public (this is especially pertinent given how widely discussed the report was in the media, see e.g. BBC 2020a, Kelly, 2020, Boseley, 2020).

In addition to assigning confidence levels to the model-outputs, the authors should have considered how model uncertainty may have impacted the quantitative values that appear in the time -series. In the first instance they could have assigned imprecise credences, representing credible intervals, to the model-outputs. In the second instance, they could have adjusted the actual values offered. As displayed in Fig. 3, these were, surprisingly, *pointed* values (rounded to two significant figures), a level of precision which is difficult to justify given the uncertainty associated with the model. In fairness, given the parameter uncertainty, particularly R_0, these could be interpreted as ranges, e.g. under the unmitigated scenario we expect 410,000–550,000 total deaths in a two year period (a range corresponding to $R_0 \in \{2, 2.2, 2.4, 2.6\}$). But these ranges are derived purely from the time- series outputs of model runs. By utilising their expert judgement, in a manner analogous to that discussed in Sect. 3.2, these ranges could have been broadened (or if the model was known to systematically under/over estimate the figures, shifted down or up respectively). And for each of these ways of communicating model- uncertainty, in order to meet **Transparent Model Uncertainty Accommodation**, and as recommended by the IPCC, the authors would be expected to provide a *transparent* (or 'traceable') discussion of their reasoning process from the model-outputs to the reported conclusions.

Moving onto the consensus report. In this instance SPI-M-O did attach likelihood modifiers to their (qualitative) conclusions (and the fact that their conclusions are offered in qualitative terms suggests that they acknowledged that precise quantitative predictions/projections were beyond their model ensemble, given the associated model- uncertainty). But, assuming that the IPCC practices are legitimate, they could have also assigned confidence judgements to these claims, judgements which would have captured their subjective evaluation of the quality of the evidence provided by each of the individual models that fed into the consensus view, the level of agreement between the models, and the method used for moving from the models to the consensus view (e.g. were some models assigned a higher epistemic warrant than others? Were systematic relationships between the models taken into account or were the models assumed to be independent? Etc.). Moreover, as noted previously, given variance in how qualitative likelihood claims are interpreted probabilistically, they could have also provided explicit guidance regarding how their language was to be interpreted. Again, meeting these recommendations would have gone some way to ensuring that **Model Uncertainty Accommodation** was met, and if it were done in a transparent (or 'traceable') way, this would have additionally met **Transparent Model Uncertainty Accommodation**. Again, this is directly motivated by the best practices developed by the IPCC.

4.2 Justifying the strategy

At this point, one might object to my line of argument along one of the following two lines.[22] First, one might worry that the analogy between the climate modelling and Covid-19 modelling is not tight enough to warrant exporting lessons from the former to the latter. Second (and independently) one might object to taking how the IPCC handle and communicate uncertainty as 'best practice' in the first place (a more general but still along these lines worry is addressed in the following section).

In response to the second concern: whilst the above discussion of transparently introducing confidence modifiers on top of quantitative probabilistic projections, and systematically adjusting said projections in light of expert judgement, are *illustrative* ways of handling and communicating model uncertainty, I am not claiming that they are the only way of doing so. The recommendations I have developed in this paper are relatively permissive, and so I grant that there are other ways of accounting for model uncertainty, and for transparently communicating how this is done to policy-makers (one of these techniques is discussed in Sect. 5.1.1), but any such way will still count as meeting my recommendations (the salient alternative is to just ignore model uncertainty altogether, by simply reporting raw model-outputs, and this is what I think was mistaken in the context of Report 9, considered in isolation). But it is worth noting that the IPCC are relative experts in this regard: their recommendations for handling uncertainty have been developed extensively through multiple iterations of their reports (Harris, 2021, provides a nice geneology of the IPCC's uncertainty framework), and whilst there is room for critical discussion, it doesn't seem unreasonable that *something like* the IPCC's framework should be seen as a best practice, and this is all that is required by my above recommendations.

[22] I'm grateful to two anonymous referees for encouraging me to explore these issues.

The concern about the tightness of the analogy between the climate and Covid-19 contexts is another matter. Two pertinent differences should be noted. First, the epidemiologists offering policy advice for responding to Covid-19 were under immediate time pressures that climate scientists authoring the IPCC reports, for better or worse, are not. As a result, one might worry that following my latter recommendation(s) for (transparently) reasoning about how to convert model-outputs to target projections would time ill-spent, given the severity of the situation. Second, the IPCC reports are written by large interdisciplinary teams, whereas the teams responsible for Report 9 and the make-up of SPI-M-O were comparatively narrower in their expertise. So the concern arises that those teams were not best placed to perform such reflection on model uncertainty.

Both of these concerns are legitimate, but I don't think they undermine the argument I am offering. With respect to time pressure: in a sense, the recommendations in question are not particularly onerous: they don't require data that were unavailable at the time, nor do they require computational power. What they require is transparent expert judgement about the limitations of the models in question. Recall that the modellers didn't develop their models from scratch: the model behind CovidSim had already been explored extensively as targeting influenza, and it is not implausible that this provided the authors of Report 9 with some understanding of its associated model uncertainty.[23] And moreover, as noted in Sect. 2.2, the SPI-M-O's consensus statement was *already* an attempt to combine multiple models in such a way as to handle the uncertainty associated with any individual model. My final recommendation is simply that the details of this aggregation should have been made transparent in (SPI-M-O, 2020g).

With respect to the second disanalogy: as noted above, I have been assuming that the modellers who construct and reason with a particular model are best placed understand the model uncertainty associated with it, and so from that perspective, one might argue that the IPCC's interdisciplinarity isn't, in fact, what contributes to their ability to handle model uncertainty. I don't think this is quite right: I suspect that the disanalogy is motivated by the thought that once a modeller has dedicated significant time and resources to a model, it is easy for them to under-recognise the potential model-target mismatches, and that by embedding such a modeller within an interdisciplinary team, the multiple backgrounds and perspectives represented can act as checks against such overconfidence in any particular model.[24] However, it is not obvious what level of interdisciplinarity collaboration is needed to ensure that policy-input reflect model uncertainty in an all- things -considered way. Moreover, Report 9 was already a highly collaborative enterprise, involving at least 31 authors, and being written on behalf of the wider Imperial Covid-19 Response team (granted, this does not entail diversity of disciplinary expertise). And, SPI-M-O itself, as already noted, involved input from *multiple* modelling teams, and the sort of robustness reasoning that, pre-

[23] Of course with the change in target system, what counts as 'model uncertainty' may also change. But in the extreme if the epidemiologists had no at hand understanding of the possible limits of their model, this raises the concern as to whether it should be used for policy-making in the first place.

[24] Thompson (2022) provides a accessible introduction to the idea that interdisciplinary can play this role. It is also related to the requirement that advisers avoid providing information laden with idiosyncratic values (Boulicault & Schroeder, 2021) an idea to which I return in the following section.

sumably featured in generating the consensus report is exactly what supports, rather than undermines, the climate/epidemiological analogy. Finally, SPI-M-O itself is supposed to reflect 'the modellers" expert consensus, which itself fed into SAGE who are ultimately responsible for communicating with policy-makers (one might worry that SAGE itself was over reliant on modellers at the expense of medics, but addressing that concern goes beyond my current scope). Taken together, I think this suffices to buttress the analogy, at least for the purposes of this paper.[25]

5 Concerns

Two assumptions underpinned the above discussion. First, that the proper role of the scientific adviser in the policy-making context is to communicate the all- things-considered state of knowledge about the relevant situation relevant for informed decision-making, and that this includes model uncertainty. Second, that doing this crucially involves drawing on the individual expertise of the scientists involved (i.e. drawing on their epistemic states, rather than the raw results of their modelling endeavours). In this section I discuss these assumptions in reverse order.

5.1 Objectivity lost

A worry with the recommendations I have been arguing for is that they introduce a *subjective* element into the process of delivering and reporting model results for use in policy-making. In a sense the information that is generated by a model is 'objective': assuming that the structure of the model is clearly documented; that the choices of parameterisation are transparent; that the random seed responsible for the stochastic behaviour of the model is publicly available; and so on, a model run, or collection thereof, is/are fully replicable by an external observer (see footnote 6). And assuming that there is an algorithmic procedure for combining a collection of models into a projection (in model land) then the resulting projection can be reconstructed from the information associated with the models, combined with that procedure.

Some of the recommendations given at the end of Sect. 4.1 do not have this 'mechanical objectivity' (*cf.* Porter, 1995; Daston & Galison, 2007, for a useful overview of objectivity in science see John, 2021). When an epidemiologist makes the decision to assign 'high', rather than 'medium', confidence in a projection, the factors that feed into this decision may not be replicable. The same applies to translating model projections into actual world projections. If an epidemiologist decides to downgrade the probability attached to an interval of individuals requiring hospital beds, from 95% in the model to 80% in the target say, the factors that feed into their decision, their knowledge of the idealised aspects of the model, may not be replicable, and nor may the way in which these factors influence the decision to downgrade the probability. So

[25] And if you are not convinced, my recommendations can be reinterpreted not as referring to the responsibilities of individual author-teams and advisory groups, but as referring to the requirement of a framework and set of best practices, for communicating model uncertainty in times where modelling endeavours are expected to feed into policy-making contexts.

one might worry that my recommendations introduce epistemically suspicious aspects into the way in which epidemiological modellers can influence policy decisions.

In this subsection I try to alleviate these worries. I first point out that there are in fact algorithmic ways of proceeding with at least two of the processes outlined above: adjusting model-outputs to projections we are more justified in imputing to their targets can be done via Bayesian techniques, which can also be utilised to combine ensembles of models to generate more justified all- thing- considered projections (as I discuss, there are other techniques available). However, I argue that there remains a need for subjective expert judgement, especially in the absence of data to which we can compare model-outputs. I then argue that this subjective expert judgement can still be legitimate by considering different senses of objectivity (Douglas, 2004).

5.1.1 Algorithmic techniques

One way of algorithmically moving from model-outputs to predictions we are more justified in imputing to their targets is described in (Kennedy & O'Hagan, 2001). The basic idea is that we can treat the 'true values' of various quantities of interest as corresponding to a combination of the model-outputs of those values and an error term that captures model- uncertainty (where each of these values is assumed to take the shape of a Guassian). This provides a 'prior' about the relationship between the model-outputs and the true values in the system. Then, we can take a sequence of model-outputs and compare them to some sequence of observed values (which may themselves contain noise). With this information we can update our prior concerning the model-target relationship to a posterior calibration of the model, and subsequent predictions about additional to be observed values.

In addition to algorithmically moving from individual model-outputs to target projections, we can also consider the process of combining model-outputs into the 'consensus view' for those projections. Here there are various different proposals on offer, which include the following: one may take a (possibly weighted) *average* of the models' results; one may apply Bayesian *stacking* techniques to the individual model results (combined with some observed data); or one may report *multiple* model results, possibly assigning each of them confidence levels, and possibly restricting which ones are reported in a manner that depends on a combination of the confidence levels and the stakes of the policy-decision they are feeding into[26]. Curiously, whilst there was no transparent discussion of how the 16 March SPI-M-O consensus view was reached (SPI-M-O, 2020g), it was later proposed that Bayesian stacking techniques be utilised to develop short-term forecasts from the models that fed into the process (SPI-M-O, 2020h, f), but as far as I can tell, this was not taken up (see, e.g. the 24 June 'SPI-M-O: Covid-19 short-term forecasts' (SPI-M-O, 2020i) where multiple model-outputs, and an equal weight average, is/are offered).

Given my current purposes, I won't delve into the details of each of these approaches beyond noting the following. First, they provide algorithmic ways in which **Transparent Model Uncertainty Accommodation** can be met, as applied to individual

[26] For a discussion of averaging vs. stacking see (Kinney 2022) for a discussion of providing multiple model results see (Roussos et al. 2021).

and ensembles of models. Second, at least some of them (particularly the ones based on Bayesian techniques) require *data* as input, data that weren't available mid-March 2020, and thus would be inapplicable during that crucial time (although could plausibly have been introduced later). Third, whilst each of these techniques looks 'objective' in the sense that they are mathematical processes taking various inputs and algorithmically generating outputs, they do at base involve subjective elements: the Bayesian techniques rely on priors; averaging techniques rely on weights assigned to models, and an equal weighting is still a subjective assessment; and the confidence based multiple model technique explicitly requires subjective confidence assignments to individual model-outputs, and subjective evaluations with respect to the appropriate confidence levels, given the stakes of the situation. So whilst using these techniques meets **Transparent Model Uncertainty Accommodation** in an at least somewhat mechanically objective way, they are only applicable under certain circumstances; they already require some level of subjective input (and are thus still subject to the concern of interest in this section); and plausibly, they fail to utilise the subjective input of experts to their full extent.

In such contexts, we are left in the situation where modellers can either pursue mechanical objectivity in their policy-advice by reporting raw model-outputs whilst ignoring the spectre of model uncertainty, or they can (transparently) account for and communicate model uncertainty, but in a manner that undermines the mechanistic nature of their advice, but hopefully, still allows for some form of objectivity. Further argument against the first disjunct is beyond my current scope (in short, my response is: modellers can chose to be mechanically inaccurate if they wish, but mechanical inaccuracy is still inaccuracy), so let's explore the second in more detail.

5.1.2 Subjective expert judgement

Return now to the idea that experts can qualify and adjust individual model projections and provide subjective inputs into the process of moving from individual models to consensus views. Recall the concern: these inputs 'taint' the process of providing objective model-based policy advice, where the latter is exemplified by the mechanical replicability of individual model runs, and some mathematical (in light of the previous discussion, still at least minimally non-mechanical) aggregation procedure. Is this the right way of thinking about objectivity in this context?

I don't think so. There are other ways of thinking about scientific objectivity that don't deliver such a verdict, and thus open the door to the sorts of techniques for meeting **Transparent Model Uncertainty Accommodation** that are recommended by the IPCC.

Douglas (2004) distinguishes between three mains forms of objectivity (each of which comes in multiple versions, thereby providing eight distinct senses of objectivity in total): *objectivity*$_1$ concerns whether we reliably track features of the world, and comes in two senses—*manipulable*, concerning how reliably we can intervene on the world, and *convergent*, concerning how the same result appears through multiple avenues—; *objectivity*$_2$ concerns individual thought processes, and comes in three senses—*detached*, according to which we shouldn't use values in place of evidence, *value free*, according to which values are banned altogether from the reasoning process, and *value neutral*, according to which the role of values in the thought process should be balanced, or neutral—; and *objectivity*$_3$, which concerns the social processes of knowledge production, and comes in three senses—*procedural*, according to which individuals involved in the process can be exchanged without impacting the end result, *concordant*, which is measured by agreement between competent observers, and *interactive*, according to which the individuals involved have to interact with oneanother when coming to their agreement.

This taxonomy can be put to work in evaluating the role (or lack thereof) of expert judgement in meeting **Transparent Model Uncertainty Accommodation**. In the first instance I do so by comparing two extremes: simply reporting model-outputs as target projections vs. utilising expert judgement to adjust these outputs when reasoning from model to target.[27]

Perhaps most importantly, in terms of manipulable objectivity$_1$, there is no clear sense in which we should expect that raw model-outputs are more accurate than the expert adjusted outputs; in fact, given that the experts are able to take into account their understanding of the uncertainties associated with their models, there is reason to think that such judgements will results in increased objectivity of this form.

The different senses of objectivity$_2$ are more interesting. It does initially seem like the expert adjustment process runs the risk of undermining detached objectivity$_2$. Allowing individual experts to adjust raw model-outputs when reporting them to policy-makers provides a route for the sociopolitical values of those individual experts to enter into the context of those reports (e.g a, agoraphobia hypochondriac may over-estimate their Covid-19 projections). And to the extent that raw model-outputs are value-free, reporting the expert adjustments rather than the raw model-outputs undermines objectivity$_2$.

There are at least two responses to this line of argument. First, with respect to the comparative claim, it is far from clear that the raw model-outputs themselves should be understood as value-free or neutral, given that they themselves depend, obviously, on decisions made when constructing the model. In general, it is well-recognised that values play a role in the construction of scientific models (see Parker and Winsberg, 2018, for a useful discussion), so the claim that raw model-outputs are independent of the values of the modellers who construct the models in question in this context would require significant defence. Second there is a reason why my third recommendation above includes the requirement that modellers *transparently* accommodate model uncertainty (or in the IPCC's terminology provide a 'traceable' account of how they have done so). By making the role of expert adjustment transparent/traceable (i.e. of the form 'the model-outputs were adjusted thus and so for such and such a reason'), the option arises that policy-makers (and others) can *evaluate* the extent of the role played by individual values in the adjustment process. These responses compound: the fact that the role of these values is 'hidden' within the construction of the model pulls in exactly the opposite direction of my urge towards transparency and the fact it allows for evaluation of the role of values in modelling for policy-making (for more on the value of transparency in releated contexts, see Elliott, 2022).

[27] Everything that I say in the next two paragraphs applies *mutatis mutandis* to the practice of offering confidence assignments on top of projections. I turn to model aggregation towards the end of this subsection.

The result of such an evaluation depends on what one requires of the role (or lack thereof) of such values in the policy-advising context. For example, according to Douglas's taxonomy, once the expert adjustment is made transparent, we can evaluate the extent to which the process is detached objectivity$_2$, i.e. we can evaluate the extent to which modellers' values have impacted their adjustment (and it is not immediately obvious that they would, model uncertainty is not intrinsically value-laden). Alternatively, if we subscribe to positions like that of Schroeder (2022), according to which scientists have a duty to present and highlight information (including model uncertainty) in a manner that is sensitive to the considered and informed values and goals of the policy-makers, then we might allow for values to play a role in the adjustment process, just so long as those values agree with the values and goals of the democratically elected policy-makers. The same applies if we want to avoid 'idiosyncratic' values from playing a role (Boulicault & Schroeder, 2021). Regardless, of the details of how this is worked out, the relevant point for my current purposes is that the transparency requirements allows us to evaluate the extent to which (if at all), and if so how, individual scientist's values have played a role in the process of providing the expert judgement on top of the raw model-outputs (which, recall, are in the first place by no means obviously value free in the first place).

We can now turn to objectivity in the social sense of objectivity$_3$. Again at first glance it seems like the practice of reporting raw model-outputs has the edge: one can exchange scientists at will without impacting the mechanically replicable outputs, which seems to tell in favour of procedural objectivity$_3$ (although, again, this is not to say that such an exchange won't impact the details of the model construction itself). But recall from the previous discussion: Report 9 was not the work of an individual modeller, it was a highly *collaborative* enterprise, involving at least 31 authors, and being written on behalf of the wider Imperial Covid-19 Response team. As a result my recommendation should be read as requiring that the authorial team *as a whole* engage in the expert adjustment of model-outputs. The implication of this is that (granting for the sake of argument that the values of the authorial team would have impacted their expert adjustments): if there are a diversity of individual values represented between the co-authors then as long as the co-authors of the report *agree* in their adjustments, and as long as this agreement is the result of an *interactive* process, then such an adjustment can legitimately said to be objective$_3$ in these senses.[28] This argument can be bolstered by considering the other target of my discussion, the development of SPI-M-O's consensus view from the outputs of multiple models (itself an example of convergent objectivity$_1$). With multiple modelling teams, the assumption that a diverse collection of values are represented becomes even more plausible. In such instances then by making the process of individual model-output adjustment, and multiple model aggregation (via any of the techniques discussed above) collaborative and transparent, we go some way to avoiding any individual 'idiosyncratic' (Boulicault & Schroeder,

[28] As Douglas (2004, p. 464) notes, what processes to use to ensure that the social process is interactive in this sense is a difficult question. See (Morgan 2014, Burgman 2016) for relevant discussions in the context of structured expert elicitation.

2021), or ‘undemocratic’ (Schroeder, 2022) values undermining the objectivity (in the relevant sense) of the policy-advice offered.[29]

5.2 The role of the scientific advisor

Finally then, a brief comment on my assumption that the role of the Covid-19 modellers, through SPI-M-O and SAGE, was to communicate their state of knowledge, sincerely taking into account model uncertainty in a transparent manner, to policymakers. Perhaps this is misguided. Perhaps, the role of scientific advisers is not to communicate information, but to prompt action (recall how time sensitive the decisions at the beginning of the pandemic were, and the ministers responsible for the decision-making).[30] And in such contexts the recommendations I have urged—adjusting, presumably by making less precise, model projections; and adding, presumably relatively weak, confidence adjusters to them—may have undermined the call to action many take to have been required at that time. Moreover, perhaps simply reporting raw model-outputs, like those displayed in Fig. 2 where the surge critical care bed capacity graph is dwarfed by the projected occupied beds, was the right thing to do.

I want to offer three responses to this concern. First, one could (although I don’t) argue that, given the all- things- considered state of knowledge at the time, nationwide, legally enforced, NPI responses were in fact unjustified, in a large part precisely because of the model uncertainty involved (for a relevant back and forth on this point see (Winsberg et al., 2020, 2021, van Basshuysen and White, 2021a, 2021b).

Second, even if it is correct that the norms governing policy-advice in the context of the early pandemic were in fact different from those governing policy-advice more generally (what Birch (2021) calls ‘science and policy in extremsis’), and the modellers in this context should have tailored their advice precisely to motivate action, it is not obvious that following my recommendations would have undermined this goal. This depends on the model uncertainties in question, and the actual state of the modellers’ all-things-considered knowledge at the time. Moreover, as Helgeson et al. (2018) note, there is a trade-off between a confidence qualifier, and the size of the interval of the first-order projection under consideration (for example, a Covid-19 modeller may be project that it is *likely* that deaths will fall in the range $[x, y]$ with *medium* confidence, or that it is *very likely* that deaths will fall within the larger range $[x - n, y + m]$ with *high* confidence). Plausibly then, the ranges that modellers were in fact highly or very highly confident in, could still have been communicated in such a way that demanded immediate action, even if they weren’t themselves model-outputs. Alternatively, reporting ‘worst-case scenarios’ deemed plausible by SPI-M-O or SAGE, informed, but not exhausted by the worst-cases in model-land, and taking into account model uncertainty in a transparent way to justify why they are so plausible, would still count as meeting **Transparent Model Uncertainty Accommodation**, and could

[29] Note here that this point is builds, but is distinct, from the discussion at the end of Sect. 4.2: there I discussed the idea that diversity of *expertise* could play a role in recognising model uncertainty, the current point is that diversity of *values* can play the role of ensuring the relevant senses of objectivity.

[30] Relevant discussions of this line of thought include (John, 2018; Birch, 2021) I’m grateful to an anonymous referee for encouraging me to consider this.

still demand immediate action. Thus, my recommendations are consistent with the all-things- considered state of knowledge about early Covid-19 being communicated in such a way that would have induced the required action.[31]

Third, at the heart of this concern lies a more general issue: in general, I submit, there is a mismatch between the standard expectations on the sort of scientific basis that is required to justify taking action in cases where action needs to be taken, and the strength of the scientific basis that can be reasonably expected from modelling endeavours in cases involve complex dynamical systems, cases involving significant model uncertainty, particularly when this is compounded by a lack of data required to get the sorts of techniques discussion in Sect. 5.1.1 off the ground. But, it's not obvious whether the appropriate reaction to this mismatch should be 'model overconfidence', as opposed to revisiting our expectations for what is needed to guide action (see Roussos et al., 2021, for relevant discussions along these lines). Whilst I cannot explore this third claim here, I hope that taken together the discussion in this subsection helps alleviates the worry that following my recommendations relies on a mistaken assumption about the role for the policy advisor in the Covid-19 context, at least to some extent.

6 Conclusion

I have argued that, at least during the crucial months during the beginning of the pandemic and particularly in the Report 9 itself, epidemiologists failed to recognise the existence of model uncertainty. They reported *just* what their models said. In doing so, they risked confusing the map with the territory. I then motivated two additional recommendations to go beyond this: they should have reported what *they* (not their models) believed, and *why* (transparently) they believed it. Drawing on the details of the IPCC's practices I suggested that experts should utilise their judgement to transparently adjust model-outputs to all- things- considered target projections, provide confidence assignments to such projections, and transparently aggregate the results of individual models to consensus views, based on model ensembles. All of which require a non-mechanical, but still appropriately objective, approach to communicating to policy-makers. Or so I have argued.

Acknowledgements Thanks to Rawad El Skaf for inviting me to participate in this project, and to two anonymous referees for outstanding feedback on this paper. I am also grateful to Roman Frigg and Margherita Harris for helpful discussions of the material during the pandemic itself. Thanks also to audiences at the Institute of Philosophy, University of London, and the Cognition, Values, Behaviour Lab, Ludwig Maximilian University of Munich.

Funding Open access funding provided by Stockholm University.

[31] See (Keohane et al., 2014) and (Schroeder, 2022) for discussions of useful comparison where the IPCC (who don't follow their own advice, i.e. (Mastrandrea et al., 2010), consistently) simply reported model-outputs that were known to underestimate projected sea levels in 2099 because the models fail to take into account a known mechanism contributing to a raise in such levels. In this instance, reporting the model-output without transparently reflecting on model uncertainty undermined policy action.

Declarations

Conflict of interest The author certifies that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

BBC. (2020a). *Coronavirus: 'act early to save more than 30 million lives'*. BBC. https://www.bbc.com/news/health-52055546

BBC. (2020b). *Coronavirus: Strict new curbs on life in UK announced by PM*. BBC. https://www.bbc.com/news/uk-52012432

Birch, J. (2021). Science and policy in extremis: The UK's initial response to covid-19. *European Journal for Philosophy of Science, 11*(3), 90.

Boseley, S. (2020). New data, new policy: Why UK's coronavirus strategy changed. *The Guardian*.

Boulicault, M., & Schroeder, S. A. (2021). Public trust in science: Exploring the idiosyncrasy-free ideal. In K. Vallier & M. Weber (Eds.), *Social trust.* Routledge.

Boulos, L., Curran, J. A., Gallant, A., Wong, H., Johnson, C., Delahunty-Pike, A., Saxinger, L., Chu, D., Comeau, J., Flynn, T., Clegg, J., & Dye, C. (2023). Effectiveness of face masks for reducing transmission of SARS-COV-2: A rapid systematic review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381*(2257), 20230133.

Bradley, R., Helgeson, C., & Hill, B. (2017). Climate change assessments: Confidence, probability, and decision. *Philosophy of Science, 84*(3), 500–522.

Burgman, M. A. (2016). *Trusting judgements: How to get the best out of experts*. Cambridge University Press.

Burki, T. (2020). England and Wales see 20,000 excess deaths in care homes. *The Lancet, 395*(10237), 1602.

Daston, L., & Galison, P. (2007). *Objectivity*. Zone Books.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese, 138*(3), 453–473.

Edeling, W., Arabnejad, H., Sinclair, R., Suleimenova, D., Gopalakrishnan, K., Bosak, B., Groen, D., Mahmood, I., Crommelin, D., & Coveney, P. V. (2021). The impact of uncertainty on predictions of the CovidSim epidemiological code. *Nature Computational Science, 1*(2), 128–135.

Eglen, S. J. (2020). Codecheck certificate 2020-010. See file LICENSE for license of the contained code. The report document codecheck.pdf is published under CC-BY 4.0 International.

Elliott, K. C. (2022). A taxonomy of transparency in science. *Canadian Journal of Philosophy, 52*(3), 342–355.

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Van Elsland, S., ... Ghani, A. (2020). *Report 9: Impact of non-pharmaceutical interventions (NPIS) to reduce covid19 mortality and healthcare demand*. Imperial College London.

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature, 437*(7056), 209–214.

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature, 442*(7101), 448–452.

Frigg, R., & Nguyen, J. (2018). The turn of the valve: Representing with material models. *European Journal for Philosophy of Science, 8*(2), 205–224.

Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*. Cham: Springer.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago University Press.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*(4), 742–752.

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese, 172*(1), 269–281.

Harris, M. (2021). Conceptualizing Uncertainty: The IPCC, Model Robustness and the Weight of Evidence. PhD thesis, London School of Economics.

Helgeson, C., Bradley, R., & Hill, B. (2018). Combining probability with qualitative degree-of-certainty metrics in assessment. *Climatic Change, 149*, 517–525.

John, S. (2018). Epistemic trust and the ethics of science communication: Against transparency, openness, sincerity and honesty. *Social Epistemology, 32*(2), 75–87.

John, S. (2021). *Objectivity in science. Elements in the philosophy of science*. Cambridge University Press.

Kelly, J. (2020). That Imperial coronavirus report, in detail. *Financial Times*. https://www.ft.com/content/1fed7551-61ce-41de-bad3-a38534b0ada8

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63*(3), 425–464.

Keohane, R. O., Lane, M., & Oppenheimer, M. (2014). The ethics of scientific communication under uncertainty. *Politics, Philosophy & Economics, 13*(4), 343–368.

Kinney, D. (2022). Why average when you can stack? Better methods for generating accurate group credences. *Philosophy of Science, 89*(4), 845–863.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science, 61*(3), 541–567.

Leung, K., & Wu, J. T. (2021). Quantifying the uncertainty of CovidSim. *Nature Computational Science, 1*(2), 98–99.

Mastrandrea, M., Field, C., Stocker, T., Edenhofer, O., Ebi, K., Frame, D., Held, H., Kriegler, E., Mach, K., Matschoss, P., Plattner, G.-K., Yohe, G., & Zwiers, F. (2010). *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC)*. http://www.ipcc.ch

Maziarz, M., & Zach, M. (2020). Agent-based modelling for SARS-COV-2 epidemic prediction and intervention assessment: A methodological appraisal. *Journal of Evaluation in Clinical Practice, 26*(5), 1352–1360.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences, 111*(20), 7176–7184.

Nguyen, J., & Frigg, R. (2022). *Scientific representation. Elements in the philosophy of science*. Cambridge University Press.

Parker, W. (2014). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science Part A, 46*, 24–30.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600.

Parker, W. S., & Winsberg, E. (2018). Values and evidence: How models make a difference. *European Journal for Philosophy of Science, 8*(1), 125–142.

Porter, T. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Razai, M. S., Kankam, H. K. N., Majeed, A., Esmail, A., & Williams, D. R. (2021). Mitigating ethnic disparities in covid-19 and beyond. *BMJ, 372*.

RC Centre for Global Infectious Disease Analysis. (2020). Covid-19 covidsim model–report 9 folder. https://github.com/mrc-ide/covid-sim/tree/master/report9

Roussos, J., Bradley, R., & Frigg, R. (2021). Making confident decisions with model ensembles. *Philosophy of Science, 88*(3), 439–460.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science, 20*(1), 1–6.

SAGE. (2022). Sage 16 minutes: Coronavirus (covid-19) response, 16 march 2020. https://www.gov.uk/government/publications/sage-minutes-coronavirus-covid-19-response-16-march-2020

Schroeder, S. A. (2022). An ethical framework for presenting scientific results to policy-makers. *Kennedy Institute of Ethics Journal, 32*(1), 33–67.

Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science, 69*(1), 275–300.

SPI-M-O. (2020a). Coronavirus disease 2019 (covid19) intervention strategies, March 16, 2020. https://www.gov.uk/government/publications/coronavirus-disease-2019-covid19-intervention-strategies

SPI-M-O. (2020b). Effect of compliance with whole household isolation in the covid-19 outbreak, March 15, 2020. https://www.gov.uk/government/publications/effect-of-compliance-with-whole-household-isolation-in-the-covid-19-outbreak-15-march-2020

SPI-M-O. (2020c). The impact of aggressively managing peak incidence, March 11, 2020. https://www.gov.uk/government/publications/the-impact-of-aggressively-managing-peak-incidence-11-march-2020

SPI-M-O. (2020d). Impact of non-pharmaceutical interventions (NPIS) to reduce covid-19 mortality and healthcare demand, March 16, 2020. https://www.gov.uk/government/publications/impact-of-non-pharmaceutical-interventions-npis-to-reduce-covid-19-mortality-and-healthcare-demand-16-march-2020

SPI-M-O. (2020e). Low critical care capacity and high severity of covid-19 mean there is little functional difference between successful curve "flattening the curve" and ongoing containment, March 16, 2020. https://www.gov.uk/government/publications/low-critical-care-capacity-and-high-severity-of-covid-19-mean-there-is-little-functional-difference-between-successful-curve-flattening-the-curve-an

SPI-M-O. (2020f). Spi-m-o: Combining covid-19 model forecast intervals, April 9, 2020. https://www.gov.uk/government/publications/spi-m-o-combining-covid-19-model-forecast-intervals-9-april-2020

SPI-M-O. (2020g). Spi-m-o: Consensus view on behavioural and social interventions, 16 March 2020. https://www.gov.uk/government/publications/spi-m-o-consensus-view-on-behavioural-and-social-interventions-16-march-2020

SPI-M-O. (2020h). Spi-m-o: Covid-19 short-term forecasting: Proposed process for discussion, April 2, 2020. https://www.gov.uk/government/publications/spi-m-o-covid-19-short-term-forecasting-proposed-process-for-discussion-2-april-2020

SPI-M-O (2020i). Spi-m-o: Covid-19 short-term forecasts, June 24, 2020. https://www.gov.uk/government/publications/spi-m-o-covid-19-short-term-forecasts-24-june-2020

Steele, K. (2012). The scientist qua policy advisor makes value judgments. *Philosophy of Science, 79*(5), 893–904.

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. M. (2013). *Climate Change 2013: The physical science basis. Working Group I contribution to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press.

Thompson, E. (2022). *Escape from model land: How mathematical models can lead us astray and what we can do about it*. John Murray Press.

Thompson, E., Frigg, R., & Helgeson, C. (2016). Expert judgment for climate change adaptation. *Philosophy of Science, 83*(5), 1110–1121.

van Basshuysen, P., & White, L. (2021a). The epistemic duties of philosophers: An addendum. *Kennedy Institute of Ethics Journal, 31*(4), 447–451.

van Basshuysen, P., & White, L. (2021b). Were lockdowns justified? A return to the facts and evidence. *Kennedy Institute of Ethics Journal, 31*(4), 405–428.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–742.

Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles—A risk for miscommunication. *Journal of Science Communication, 19*(02), A03.

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. *Kennedy Institute of Ethics Journal, 22*(2), 111–137.

Winsberg, E., Brennan, J., & Surprenant, C. (2021). This paper attacks a strawman but the strawman wins: A reply to Van Basshuysen and white. *Kennedy Institute of Ethics Journal, 31*(4), 429–446.

Winsberg, E., Brennan, J., & Surprenant, C. W. (2020). How government leaders violated their epistemic duties during the SARS-COV-2 crisis. *Kennedy Institute of Ethics Journal, 30*(3), 215–242.