Insights From Insurance for Fair Machine Learning

Christian Fröhlich *christian.froehlich@uni-tuebingen.de*
Department of Computer Science
University of Tübingen
and Tübingen AI Center

Robert C. Williamson *bob.williamson@uni-tuebingen.de*
Department of Computer Science
University of Tübingen
and Tübingen AI Center

Abstract

We argue that insurance can act as an analogon for the social situatedness of machine learning systems, hence allowing machine learning scholars to take insights from the rich and interdisciplinary insurance literature. Tracing the interaction of uncertainty, fairness and responsibility in insurance provides a fresh perspective on fairness in machine learning. We link insurance fairness conceptions to their machine learning relatives, and use this bridge to problematize fairness as calibration. In this process, we bring to the forefront two themes that have been largely overlooked in the machine learning literature: responsibility and aggregate-individual tensions.

1 Introduction

Insurance is "interestingly uninteresting".[1] In this work, we argue that in fact insurance is far from uninteresting and indeed a rich source of inspiration and insight to scholarship interested in social issues surrounding machine learning, specifically the field now known as fair machine learning. Our proposal is that insurance can be viewed as an analogon to machine learning with respect to these issues arising from the social situatedness. While machine learning is a relatively recent technology, debates regarding social issues in the context of insurance have been ongoing for a long time. Thus, we argue that taking inspiration from studies of insurance can contribute to a more integrative view of machine learning systems as *socio*-technical systems (Selbst et al., 2019).

Both machine learning and insurance are firmly based on a statistical, probabilistic mode of reasoning — an *actuarial* mode. Indeed, insurance can be viewed as the first commercial test of probability theory (Gigerenzer et al., 1989; McFall, 2011). Insurance, a technology for *doing risk*, transforms uncertainty into calculable risk (Lehtonen & Van Hoyweghen, 2014). The key idea is to share the risk of a loss in a collective, organized through an abstract mutuality; due to the 'law' of large numbers, uncertainty thus becomes manageable and the effect of chance can be offset (Ewald, 1991). In this way, insurance creates a "community of fate" in the face of uncertainty (Heimer, 1985). To enter into this community (the insurance *pool*), the insurer demands a certain fee, called *premium*, from the policyholder.

In insurance, questions of fairness inevitably arise, and have been the subject of much debate. The central point of debate is the tension between risk assessment and distribution (Abraham, 1985). In other words, who is to be mutualized in the pool. Some form of segmentation is found in many insurantial arrangements: the pool of policyholders can be stratified by separating high and low risk individuals. But the specific nature that such segmentation

[1] McFall et al. (2020) call insurance "interestingly uninteresting", referring to how insurance is "hugely underresearched" given its societal importance, which is typically not recognized (Ewald, 1989).

takes typically depends not only on risk assessment, but on further considerations such as assignment of responsibility, modulated by social context; in this way, insurance is not a neutral technology (Baker & Simon, 2002; Glenn, 2003a).

Our non-comprehensive outline of the history of insurance illustrates how uncertainty, fairness and responsibility interact, and can be entangled and disentangled. From this background, we can extract conceptual insights which also apply to machine learning. The tension between risk assessment and distribution is mirrored in formal fairness principles: *solidarity*, which can be linked to independence in fair machine learning, contrasts with *actuarial fairness*, linked to calibration. Briefly, actuarial fairness demands that each policyholder should pay only for their own risk, that is, mutualization should occur only between individuals with the same 'true' risk. In contrast, solidarity calls for equal contribution to the pool. On one level of this text, we problematize actuarial fairness (by extension, calibration) as a notion of fairness in the normative sense by taking inspiration from insurance. This perspective is aligned with recent proposals that stress the discrepancy of formal algorithmic fairness and "substantive" fairness (Green, 2022), which some prefer to call *justice* (Vredenburgh, 2022). Parallel to this runs a distinct textual level, where we emphasize two intricately interacting themes: *responsibility* and *tensions between aggregate and individual*. Both entail criticism of actuarial fairness, but we suggest that they additionally provide much broader, fruitful lessons for machine learning from insurance.

At the highest level of abstraction, our goal is to establish a general conceptual bridge between insurance and machine learning. Traversing this bridge, machine learning scholars can obtain new perspectives on the social situatedness of a probabilistic, statistical technology — we attempt to offer a new 'cognitive toolkit' for thinking about the social situatedness of machine learning. Our point of view is that fairness cannot be reduced to a formal, mathematical issue, but that it requires taking broader social context into account, reasoning for instance about responsibility. And for this, we suggest, insurance is an insightful analogon. Therefore, our objective is to furnish the reader with a guide that charts the landscape of insurance with respect to social issues and to establish links to machine learning.

On a formal level, we use the following analogy. In a machine learning task, we are given some features X and associated outcomes Y, which we attempt to approximate by predictions $\hat{Y}$. The structural relation to insurance is established by conceiving of X as the features of policyholders (e.g. age, gender) with outcomes Y (e.g. having an accident or not), and the task is to set a corresponding premium $\hat{Y}$.

2 A Brief History of Insurance Rationalities

Insurance is not a monolithic technology, but rather a general principle of risk management, which is instantiated in multiple distinct forms. Insurance and the conceptual resources it deploys are not immutable and stable over time, as exemplified by Baker's (1996) study on *moral hazard*, attesting to their evolving nature. In this section, we provide a succinct, necessarily non-comprehensive description of three historical modes of insurantial operation: the welfare state, neoclassical economics and personalized insurance. Throughout, we focus on the role that uncertainty, fairness and responsibility play in each of the three modes. To each mode we ascribe a set of attitudes towards these three aspects and in what manner they are entangled or disentangled. Fairness conceptions in insurance are contingent upon prevailing societal norms, particularly regarding responsibility, but concurrently insurance shapes the moral fabric of the society in which it is embedded (Glenn, 2003b; Van Hoyweghen et al., 2006; Lehtonen & Liukko, 2015). Furthermore, fairness conceptions in insurance are historically intertwined with their accompanying (statistical) epistemologies, ways of 'knowing' the risk in the face of uncertainty. A common thread is also that distinct forms of insurance correspond to distinct ways of governing society. Importantly, we do not want to suggest a linear historic progression here[2] — different forms of insurance co-exist at any given time. For instance, contemporary health care systems tend to operate with the logic of the welfare state, while actuarial fairness undergirds the private insurance sector. What follows is our synthesis of the literature, particularly drawing

[2] See Baker (2000, p. 571) on Ewald's work for this point.

on the works of Ewald (1986) and Frezal & Barry (2020), with a discerning focus on elucidating the intricate interplay between uncertainty, fairness, and responsibility. With this background, we are then able to extract conceptual lessons that apply also to machine learning.

2.1 Broad Solidarity in the Welfare State

In his seminal work *L'Ètat providence*, Ewald (1986) gave an influential account of the rise of the welfare state and explicates how insurance became its prime way of government. The point of departure is the predominance of liberal reasoning, which operates with the categories of *fault* and *foresight* in risk management. Here, it is presumed that the occurrence of an accident (a damage, loss or injury) must be due to fault, a lack of foresight. Liberal thought held individuals responsible for their own fate and inequalities were naturalized as just consequences of individual responsibility (Landes, 2013). An accident implies a trial under the "regime of juridical responsibility" (Ewald, 1986), where the goal is trying to establish the fault of one party. Responsibility is then borne by the person who is assigned fault, who is seen as having caused the accident. In turn, voluntary charity is the preferred means of supporting the poor and unlucky. With the rise of industrialization in different countries, an "accident crisis" unfolded roughly between 1870 and 1910 (Krippner, 2023): the number of workplace accidents dramatically increased and there appeared a new regularity at the aggregate level, suggesting a kind of determinism in the phenomenon. This led to the *objectification* of the accident and its management became a question of the collective, the *social*. By objectification, we understand an act of aggregation combined with the law of large numbers. As a consequence, we find a conception of insurance as broad solidarity and the rise of the welfare states. While at the individual level it was impossible to predict *who* will suffer an accident, the new regularity observed at the aggregate level could provide an effective pattern of risk management. Since equal ignorance in the fate of uncertainty was emphasized (Ewald, 1986), solidarity, implying equal contribution to the pool, was considered fair. Indeed, this conception of insurance was so firmly based on the aggregate that it led Ewald (1991) to assert that

> Strictly speaking there is no such thing as an individual risk; otherwise insurance would be no more than a wager. Risk only becomes something calculable when it is spread over a population. The work of the insurer is, precisely, to constitute that population by selecting and dividing risks. [..] It makes each person a part of the whole.

The business of insurance, then, was construed as the constitution of abstract mutualities[3] (Lehtonen & Liukko, 2011) and the sharing of responsibility to counteract the effect of fate.

The epistemology of the welfare state is one of the aggregate, the collective, the social. In this respect, it is interesting and instructive how insurance became intertwined with probability and statistics. Early forms of insurance were more like *gambling*, and insurance was often accused of being immoral and faced prohibitions.[4] Insurance gained more legitimacy when it became based on 'objective' probability and statistics (Daston, 2023, p. 162ff); by the end of the nineteenth century, the morality of insurance was established (Baker, 1996). The crucial conceptual move was the marriage of probability theory and statistics. While early probabilists were more concerned with reasonable subjective judgment, the nineteenth century shows an increasing shift towards "objective calculation" (McFall, 2011). The idea was that in a large insurance pool losses occur randomly, so the law of large numbers applies and the total loss can be predicted at the level of the aggregate. Frequentist ('objective') probability thus combines aggregate regularity with individual irregularity, as explained by Venn (1876, p. 4).

[3] Typically policyholders are unaware with whom they are mutualized, however, so this mutualization does not require a shared sense of groupness (Krippner & Hirschman, 2022).

[4] Cooper & Grinder (2009) provide some examples. In 18th century London life insurance could be bought on the life of celebrities, without an insurable interest to the policyholder. In fact, the notion of an insurable interest was put forward by the insurers to counter the allegation of gambling (Baker, 1996; McFall & Moor, 2018). Particularly interesting is also insurance in Islamic law, which prohibits gambling and contracts based on usury: the morality of insurance is justified then by emphasizing the solidaristic nature of the arrangement (Baker, 2002), in contrast to the view of insurance as a bilateral contract that is more prevalent in Western societies.

A key player in this development was Adolphe Quetelet, a Belgian astronomer, who initiated the study of "social physics" by attempting to discover natural laws about human behaviour (McFall, 2011). Quetelet's innovation was the transposition of one sense of the concept *average* to a different one. Consider first the familiar aggregating sense of average, where a set of commensurate objects is summarized in a single number. A prima facie different sense of average is as the single true value of some measurement problem, from which one can obtain a set of noisy measurements. The radical conceptual move was then to transpose the second to the first sense, thereby viewing the individuals of a population as many realizations of some abstract *average human*[5] (Ewald, 1990; McFall, 2011). In this way, the rates of birth, death and other social phenomena could be attributed to this fictional average human. In the context of the workplace accident and insurance, in line with the imaginary of the average human, the focus shifted from the unique, individual experience (the object of a juridical trial) to an objectification based on the average occurrence (Krippner, 2023). With regard to how this influences the notion of personhood, Dean (1998) writes

> Insurance practices displace the abstract, invariant norm of a responsible juridical subject with an individuality relative to other members of an insured population, an 'average sociological individuality'.

Quetelet's fiction of the average human has made an impact on the insurance sector (McFall, 2011): identifying the individual with an average is at the core of actuarial practice. Indeed, the question of how individuals relate to the aggregates they make up is, as we suggest in line with Krippner (2023), runs through the history of insurance. Moreover, we argue in Section 5 that it is a major concern for fair machine learning, too.

Although insurance increasingly relied on probability and statistics, the available quantification methods at the level of the collective severely limited the possibility of actually 'knowing' the risk of an individual (Barry, 2019). With improved actuarial methods, new possibilities for segmentation of the pool have opened up, which have led also to the rise of a new fairness notion that is successively contributing to the erosion of solidarity.

2.2 Neoclassical Economics and Actuarial Fairness

A distinct mode of insurance, which undergirds contemporary (private) insurance, is based on neo-classical economics, which construes individuals as rational expected utility maximizers — the human is viewed as a *homo oeconomicus*. The assumption in this paradigm is that insurance is purchased due to risk aversity from the perspective of the policyholder, while the insurer is risk neutral. This configuration of the individual is tied to a different notion of fairness which contrasts starkly with the solidarity of the welfare state: *actuarial fairness*. The idea is that the pure premium (i.e. what the policyholder pays before adding additional expenses such as for administration) should equal the expected risk for each policyholder. While the idea of "equality in risk" was around for a long time (Heras et al., 2020), its modern formulation is due to Arrow (1963). On the one hand, actuarial fairness can be understood as a purely descriptive, technical notion; but it is also advanced in the literature and by insurers as a notion of *fairness* in a normative sense, as legitimate practice, see for instance (Walters, 1981; Clifford & Iuculano, 1987; Daniels, 1990; Stone, 2001; Thiery & Van Schoubroeck, 2006); Interestingly, actuarial fairness can be traced back to the Aristotelian consistency principle "fairness is to treat equal people equally and unequal people unequally" (Landes, 2015), in a context where 'equal' means 'equal risk' (Heras et al., 2020).

The definition and especially the Aristotelian motivation betrays that actuarial fairness is in practice always a *group-based* notion (Miller, 2009), since insurers needed (traditionally, at least) to make use of large segments for calculating expected losses. For this, actuaries choose a set of relevant variables, while ignoring others. This seems fundamentally in conflict with the idea of adjustment to the 'individual risk' of the policyholder — and indeed this

[5] The original term was "average man".

was not how actuarial fairness was construed until roughly the 1970s, it remained firmly group-based (Barry, 2020). In line with Quetelet, Thiery & Van Schoubroeck (2006) describe the logic succinctly as follows:

> [I]nsurance classification schemes rely on the assumption that individuals answer to the average (stereotypical) characteristics of a group to which they belong.

Hence the justification for group-based actuarial fairness relies on assuming an "average sociological individuality" (Dean, 1998) — each individual is assigned to a segment and is then identified with the corresponding average (Krippner & Hirschman, 2022). In this sense, we find again the logic of the welfare state but now only *segmentwise*, with the aspiration to reduce solidarity between groups as much as possible, given practical constraints. In the terminology of Lehtonen & Liukko (2011), this means that ideally only *chance solidarity* is left, which compensates for the effects of aleatoric uncertainty;[6] in contrast, *risk subsidizing solidarity* refers to a solidarity between individuals of different expected loss.[7]

The role of responsibility in actuarial fairness is subtle. Actuarial fairness, when understood as a normative principle, rests on the assumption that people can be held responsible for their individual risk to some extent (Lehtonen & Liukko, 2015). However, we should distinguish conceptually between *responsibility* and *responsibilization*, that is, *holding* someone responsible for something.[8] While normative philosophical literature may be careful about this distinction, in practice it can appear blurry. In fact, there are two principles that provide *non-responsibility based reasons for responsibilization* in insurance (Andersen & Nielsen, 2015), therefore in favor of actuarial fairness: *moral hazard* and *adverse selection* (see Appendix B). Thus, responsibility is, in the neoclassical framework, not central to actuarial fairness (Landes, 2015). However, the role of responsibility (more precisely, responsibility-based reasons for responsibilization) in insurance is currently being emphasized more and more. To this we turn now.

2.3 The Climax: Personalized Insurance

From the 1980s on, the insurance industry was increasingly challenged by anti-discrimination legislation. Social movements attacked the average human (woman) logic that insurance based its actuarially 'fair' premia on. A prominent example is the campaign initiated by the National Organization for Women (NOW) (Krippner & Hirschman, 2022; Krippner, 2023), aimed at ending gender-based risk segmentation. In line with the civil rights movement, feminists considered such underwriting (that is, risk classification) practices to be unfairly discriminatory, as they rely on group-based generalizations. Instead, they asked for a finer adjustment to *individual risk*. What was under attack here is the fundamental group-based logic of actuarial fairness. The US supreme court asserted in the context of insurance that "[e]ven a true generalization about [a] class cannot justify class-based treatment" (Norris, 1983, as cited in Avraham (2018)). A more recent case is the *ECJ Test-Achats* ruling in the EU, which highly restricts gender-based underwriting (Rebert & Van Hoyweghen, 2015; Cevolini & Esposito, 2022). In response to such anti-discrimination legislation and in anticipation of a continuation of this trend, the practical meaning of actuarial fairness gradually began to shift. Consequently we find two highly entangled trends in the contemporary insurance industry: the *individualization of risk* and *behaviour-based personalization* (Cevolini & Esposito, 2020; McFall et al., 2020).

The individualization of risk can be understood as taking group-based actuarial fairness to the limit and (hypothetically) forming 'groups of one'; instead of spreading risk over a pool, each policyholder would pay exactly for her own

[6] We put aside for now the question of whether the conceptual distinction between aleatoric and epistemic uncertainty is well-defined.

[7] Lehtonen & Liukko (2011) further mention *subsidizing income solidarity*, occuring when premia are adjusted based on income; this is more like a tax than 'genuine' insurance solidarity.

[8] We use the term 'responsibilization' in line with Andersen & Nielsen (2015). The term originally appeared in the governmentality literature and refers to a neoliberal mode of governing that frames individuals as autonomous and responsible, see for example (Shamir, 2008; Pyysiäinen et al., 2017).

individual risk (Cevolini & Esposito, 2020). In terms of solidarity, this implies a complete erosion of subsidizing risk solidarity, so that ideally only chance solidarity remains. Rephrasing this, what is now being emphasized is the non-homogeneity within previous groups of policyholders (Barry, 2020). To this end, insurers begin to shift the focus from attributes which are considered uncontrollable (e.g. gender, race)[9] to controllable, dynamic data about the individual, and adjust premia accordingly, yielding behaviour-based personalization. Of course, the hope is that behaviour is closely linked to the individual risk, otherwise personalization would hardly be reasonable; in this way individualization and personalization are correlated.[10]

Personalization is linked to *InsurTech*, that is, technology-driven innovation in insurance (McFall et al., 2020). A prominent example is the use of wearable devices such as fitness trackers in health insurance (Lupton, 2016; McFall, 2019), where discounts and rewards are supposed to incentivize "healthy behaviour". In car insurance, the use of telematics is gaining popularity (Verbelen et al., 2018; Meyers & Van Hoyweghen, 2018; Cevolini & Esposito, 2022), where a small device installed in the car dynamically provides information about driving behaviour from proxy variables such as speed to the insurer, as well as feedback to the policyholder. Here, the premium is continuously adjusted to behaviour, which contrasts with previous static underwriting. In both examples, the premium is supposed to act *on* the behaviour with the aim of loss prevention (McFall & Moor, 2018); this opens up the possibility for feedback loops. In other words, the premium is *performative* — see Appendix D.

The accompanying epistemology associated with the shift towards individualization and personalization is the aim to tailor the premium to the 'individual risk' and it is assumed that a combination of big data (often behavioural, with a fine temporal resolution) and machine learning enables 'knowing' this risk (Cevolini & Esposito, 2020). Hence, in this currently unfolding chapter, machine learning enters into a dynamic interaction with insurance; we expect conceptual lessons to flow in both directions in the future (compare also (Williamson, 2004)); in this paper, however, we specifically focus on lessons from insurance for machine learning. Barry (2019) describes the new epistemology as follows:

> Hence what was once considered as 'noise,' the individual specificities that had to be averaged out by statistics, is now the core of the analysis and the focus of the new knowledge.

The upshot, according to Barry (2019), is the "deconstruction of the aggregate viewpoint that produced collectives". Commentators speak of emerging "segments of one" (Prainsack & Van Hoyweghen, 2020).

The individualization and personalization of risk is associated with a shift in fairness: actuarial fairness is taken to the limit and now clearly carries a normative flavour based on a linkage to *responsibility*. We call this utopia of individual risk adjustment *perfect actuarial fairness*, to demarcate it from practical, group-based actuarial fairness. Reviving pre-Welfare liberal thought, individual responsibility is stressed (Dean, 1998). For example, Ericson et al. (2000) document a shift in the concept of accident; the new rhetoric, speaking of a "crash" in the case of a car accident, underscores that *someone must be at fault* and thus responsible. Without taking a philosophical stance on actual responsibility, the trend is one of the *responsibilization* of the individual. Even in the context of health insurance, individuals face such responsibilization (Van Hoyweghen et al., 2006; Prainsack & Van Hoyweghen, 2020). For example, self-tracking favors a view of individuals as "managers" of their health (Lupton, 2016; Sharon, 2017). Overall, responsibilization is linked to neoliberal modes of government (Dean, 1998; Ericson et al., 2000; Meyers & Van Hoyweghen, 2018).

As a consequence, many commentators argue that personalization undermines solidarity (Rosanvallon, 2000; Prainsack & Van Hoyweghen, 2020; Barry, 2020; Cevolini & Esposito, 2020); For instance, Swedloff (2014) claims that big data is in contradiction to the risk-spreading mechanism of insurance and Heras et al. (2020) observe that,

[9] While the case of gender demonstrates that what is considered controllable can change, for insurance purposes, gender is arguably still uncontrollable.

[10] One would (in most contexts) not try to personalize premia based on the binary feature 'having attached earlobes'.

when taken to the extreme, actuarial fairness contradicts the very logic of insurance.[11] When risk spreading disappears, insurance becomes more like personal saving. With respect to distributive consequences, the individualization of risk has been "profoundly inegalitarian" (Armstrong, 2005). The highest-at-risk individuals can even face exclusion from the pool (Lehtonen & Liukko, 2015; Cevolini & Esposito, 2020).

In summary, we have described three broad modes of insurance and their associated attitudes towards uncertainty, fairness and responsibility. We now investigate multiple dimensions of responsibility and then establish a link to fair machine learning, where we argue that reflections on responsibility should be foregrounded.

3 Responsibility

Insurance actively constructs and distributes responsibility (Baker, 2002); the boundaries of individual responsibility are drawn by accounting for some factors but not for others in the premium. Adding to the terminology of Landes & Holtug (2015), we distinguish four dimensions of responsibility: causal (*who is causally responsible for the accident?*), control-based (*who could control the happening?*), moral (*who is normatively responsible?*) and material (*who bears the consequences?*). By responsibilization we understand holding individuals responsible, with an emphasis on the material dimension, based on narratives about causal, control-based and moral dimensions of some phenomenon. What is distinctive about insurance as a technology of risk management is the tendency to separate these dimensions (Landes & Holtug, 2015). However, we have observed a recent trend towards a renewed entanglement when compared to the mode of the welfare state. In particular, the notion of actuarial fairness has been increasingly linked to responsibility in contrast to mere non-responsibility based responsibilization.

What is intriguing is also how the entanglement of uncertainty and responsibility has changed historically. A thick veil of ignorance, when the individual level remains out of reach, appears to favor a collective responsibility for risk management; when this veil is gradually lifted, it seems easier to assign responsibility to individuals (Frezal & Barry, 2020; Barry, 2020). However, this is not a necessity: 'knowing' individual risk (to some extent) does not necessarily imply that individuals are morally responsible or that we should hold them materially responsible (see Section 3.2).

3.1 Causality and Control

Causality has received major attention in recent machine learning research and is also widely discussed in the literature on insurance. A highly related, but subtly distinct notion is that of control — indeed, the recent trend of responsibilization can to some extent be explained as a response to legislation that prohibits differentiating premia based on variables beyond individual control (Meyers & Van Hoyweghen, 2018). While actuarial practice is correlation-based, causality and control are emphasized by legal commentators on insurance discrimination (see e.g. (Abraham, 1985; Gaulding, 1994; Avraham et al., 2013; Avraham, 2018)). For example, Avraham (2018) demands that a variable used for calculating premia must be both causally linked to the outcome of interest *and* within individual control. Why demand a causal relationship? If a variable is merely non-causally correlated with the outcome, then it is a proxy for another variable, which should be used in its place in order to avoid differential inaccuracy (Avraham et al., 2013).

Another argument is that the importance of causality is derived from control, and that it is control which is at the heart of many controversies. To fix a rough notion of causality, we understand "X causes Y" as "intervening on X changes the probability for Y", where an intervention means changing the value. In this way, we can view causality as hypothetical control. Further, if an individual can *actually* intervene on X, then we may say simply that X is

[11] Some argue that there is no place for risk subsidizing solidarity in insurance, that insurance is concerned only with chance solidarity. However, the conceptual distinction between these forms of solidarity is unclear and rather heuristic (Frezal & Barry, 2020); cf. also the discussion in Section 5 on individual risk.

under control of the individual.[12] Observe that "X causes Y" is thus a necessary but not sufficient condition for an individual's ability to control Y by controlling X. The importance of control in turn derives from responsibility: how could an individual be responsible (in the moral and perhaps in the material sense) for a variable which is beyond control (Abraham, 1985; Avraham, 2018)? Mere causality seems insufficient for this. Hence, we view causality as the conceptual entry point to get at control. Moreover, control is key for downstream effects, as we discuss in Appendix D. From a normative perspective on responsibility, the importance of control (or more precisely, choice) is emphasized in theories of *luck egalitarianism*, often discussed as a justification of risk classification (Knight, 2013; Lippert-Rasmussen, 2015; Huseby, 2016; Björk et al., 2020).

Conversely, control is arguably not *sufficient* for responsibility. For instance, control without causality can yield 'discrimination by proxy': in Swedloff's (2014) hypothetical example, liking Vampire novels (arguably within control) is correlated with risky behaviour, but in a non-causal way. This suggests that here a controllable variable works as a proxy for a potentially non-controllable one such as gender. To further complicate matters, the argument by Hu & Kohler-Hausmann (2020) demonstrates that a conflict may arise when a seemingly controllable stands to a non-controllable variable such as gender in a *constitutive* relation;[13] responsibilizing for the controllable variable then effectively leads to responsibilization for the non-controllable one, too.

In the context of (fair) machine learning (see Section 4), causality has received much attention, but due to the previous considerations we suggest putting reflections on control-based responsibility at the center. This also implies shifting the focus to downstream ('performative') effects of deploying a machine learning system, since consequences of responsibilization are linked to control (see Appendix D). Problematically, however, we must answer the question of what is under control. We now argue, in line with other social studies of insurance scholars (Abraham, 1985; Gaulding, 1994), that the variant of this question which is relevant for insurance and machine learning purposes is in fact fundamentally *normative* in character.

3.2 Social Contingencies in Responsibilization: What Is Under Control?

The distinction of *control vs. no-control* plays a major role in justifying the responsibilization of individuals, and thus actuarial fairness in the mode of personalization.[14] Here we illustrate with examples from insurance that this dichotomy is a slippery one, however, and thus provides only a shaky basis for actuarial fairness. The question of whether a variable is within individual control is often unclear and in fact insurance scholars have argued that it is a normative question (Abraham, 1985; Gaulding, 1994), not a descriptive one. Our sketch of an argument is as follows.

An instructive example is risk classification based on smoking in life insurance. It was only in the 1980s that life insurance companies widely started to differentiate premia with respect to smoking, even though it was already known in the 1960s that the associated difference in lifespan is substantial — the decision *not* to responsibilize depended on the social acceptability of smoking (Wilkie, 1997; Glenn, 2003b). Today, 'lifestyle'-based responsibilization increasingly plays a role in insurance and smoking is considered a prime example for a variable within individual control (Van Hoyweghen et al., 2006; 2007). We should thus pay close attention to shifting societal narratives about control and responsibility. A contrasting example is the case of HIV in underwriting practices, which Daniels (1990) explicitly contrasts with the (previous) neglect of smoking as a rating variable. Daniels (1990) discusses a widespread denial of health insurance coverage for individuals with HIV with a justification based on actuarial fairness. However, Daniels (1990) suspects that homophobia and social antipathy to drug users may play a role in this, favoring responsibilization for risky 'lifestyle', conceived as controllable.

[12] As an example, genes might be causally related to some outcome Y of interest (hypothetically controlling genes would change the probability for Y), but are not actually under control of the individual.

[13] For constitutive relations, see also the discussion on performativity in Appendix D.

[14] Recall that in the neoclassical mode, the question of control is disregarded, but it is emphasized in the mode of personalization.

Underwriting based on 'lifestyle' risks can be most starkly contrasted with the use of genetic information in health and life insurance, which is now tightly regulated in some countries (Van Hoyweghen, 2018) albeit welcomed by the industry (Rechfeld, 2016). The notion that "we are all carriers of genes" has successfully invoked a solidaristic imaginary in this context (Van Hoyweghen et al., 2007): many argue that nobody should be penalized for their genes, since they are clearly beyond individual control. Here, a 'genetic veil of ignorance' is mobilized in the debate. This has even given rise to the paradigm of *genetic exceptionalism*, holding that genetic information is normatively distinct from other medical information (for a critique see (Lippert-Rasmussen, 2015)). In the context of genetic information, solidarity is thus emphasized (Liukko, 2010), but this has on the other hand contributed to a responsibilization of 'lifestyle' risk which continues to justify actuarial fairness (Van Hoyweghen, 2018).

Do the previous cases really show that the question of control is a normative one? We are not opposed to the idea that there exists a prior, descriptive question of control: considering *all* possible actions[15] that an individual can embark on, does one of them intervene on X? In this way, it seems reasonable to say that for instance driving behaviour, but not genes, are controllable. This, however, misses the point as it is not the relevant question in the context of responsibility. Control-relevant actions will bring about different consequences for the individual, and what is more, those consequences will differ among individuals. To give up smoking might give more negative utility to an individual with genetic dispositions that favor addictive behaviour. To move to a less earthquake-prone area, in order to lower one's insurance premium, might require investing a large amount of one's resources. Reasoning about the control dimension of responsibility then amounts to setting the boundaries of which actions we may justifiably demand from an individual, and hence becomes normative in character, in effect a matter of distributive justice. Acknowledging the normative element in questions of control offers a new lens on the *fairness* of actuarial fairness, demonstrating that actuarial calculations are not as 'objective and neutral' as they are promoted, a point which we further develop in Section D. While our examples are from insurance, we want to transport conceptual insights to machine learning.

4 The Link to Fair Machine Learning

In recent years, as machine learning is increasingly being deployed in sensitive domains, the field of *fair machine learning* has flourished (Barocas et al., 2019; Mitchell et al., 2021; Mehrabi et al., 2021; Castelnovo et al., 2022). Different mathematical formalizations of fairness have been proposed in the literature (see e.g. (Barocas et al., 2019)); we here focus on statistical, group-based definitions. As is common in the literature we fix a probability space, for simplicity assumed finite, where we define the following random variables: $X \in \mathcal{X}$ represents the features of the individuals under consideration, Y represents the true outcomes associated with those individuals, $\hat{Y}$ represents the predictions generated by our model, and $S \in \mathcal{S}$ represents a 'sensitive feature'[16] related to the individuals. We assume that S can be perfectly predicted from X, e.g. $X = (\tilde{X}, S)$ for some $\tilde{X}$. For simplicity, we assume binary $Y \in \{0, 1\}$ and probabilistic scores $\hat{Y} \in [0, 1]$. For example, in a credit lending scenario, X contains features such as age, income etc., S could represent "having migrant background" in a binary way, Y indicates whether an individual defaulted or not and $\hat{Y} \in [0, 1]$ represents the probabilistic prediction of the model. Group-fairness definitions are often based on binary decisions, but for the analogy with insurance we use probabilistic scores: we intuitively think of Y as representing the true outcome (an accident, damage or loss) and $\hat{Y}$ as representing the premium that the insurer (by analogy, the ML engineer) demands for shouldering the risk of the uncertain outcome Y. By $\perp\!\!\!\perp$ we denote statistical independence of random variables.

Definition 4.1. *A model satisfies* independence *if* $\hat{Y} \perp\!\!\!\perp S$.

If, for instance, S represents migrant background, independence demands that the distribution of scores is the same for people with and without migrant background. Independence starkly contrasts with calibration.

[15] Even granting that such an expression is sensible.

[16] In the fair machine learning literature, a sensitive feature relates to membership in a socially salient group, for instance based on gender, race or religion.

Definition 4.2. *A model satisfies* calibration by groups *with respect to S if*

$$\mathbb{E}[Y \mid S = s, \hat{Y} = \hat{y}] = \hat{y}, \quad \forall s \in \mathcal{S}\ \forall \hat{y} \in [0, 1].$$

As a fairness criterion, calibration embodies the aim of matching our probabilistic predictions well to the true outcomes. If our model predicts a probability of p% for defaulting, then indeed p% should default if our model is adequate. Recently, Höltgen & Williamson (2023) have demonstrated that in fact calibration is a richer notion than what is captured by the traditional definition. While they focus on the case of finite data, we present a corresponding theoretical definition. Assume again some set of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. Calibration can then be defined based on this choice of groups.

Definition 4.3. *A model satisfies (theoretical) calibration with respect to a set of groups $\mathcal{G}$ if:*

$$\forall G \in \mathcal{G} : \mathbb{E}[(\hat{Y}(X) - Y)|X \in G] = 0.$$

While this looks deceptively similar to the recently popularized *multi-calibration*, it abstracts away from the prediction-based binning, which is partly due to historical reasons (Höltgen & Williamson, 2023). To gain intuition, it is instructive to consider what this means in the case of finite data. Assume a finite dataset $(X_1, Y_1), .., (X_n, Y_n)$ with associated predictions $\hat{Y}(X_1), .., \hat{Y}(X_n)$. A set of groups is then equivalent to choosing a subset of the data, i.e. a set $\mathcal{G} \subseteq 2^{\{1,..,n\}}$. If we use the empirical distribution associated with this dataset in Definition 4.3, we obtain the following empirical variant.

Definition 4.4. *A model satisfies (empirical) calibration with respect to a set of groups $\mathcal{G}$ if:*

$$\forall G \in \mathcal{G} : \frac{1}{|G|} \sum_{i \in G} (\hat{Y}(X_i) - Y_i) = 0.$$

This offers an insurantial interpretation: $\hat{Y}(X_i)$ is the premium we demand for shouldering the risk of the uncertain Y_i. Indeed, calibration is formally reminiscent of the subjective betting interpretation for probability theory proposed by de Finetti (1974/2017). There, the expectation $\mathbb{E}[Y]$ is viewed as the fair betting price for a gamble Y.[17] Calibration is however not tied to a subjective or objective interpretation of probability, but a criterion for model evaluation.

Actuarial fairness and calibration are in close correspondence. For a given set of groups $\mathcal{G} \subset 2^{\mathcal{X}}$, which we assume forms a partition of $\mathcal{X}$, we define the actuarially fair predictor $\hat{Y}_{\text{af}}(x) \coloneqq \mathbb{E}[Y|X \in G_x]$, where G_x is the unique group that contains x. The goal of actuarial fairness is to make this partition as fine as possible (cf. Section 2.2, 2.3).

Consider first the extreme choice of a single group as the whole population in Definition 4.4. Then calibration in the theoretical (respectively, empirical) case demands that

$$\mathbb{E}[\hat{Y}(X) - Y] = 0, \quad \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}(X_i) - Y_i) = 0,$$

which is called *global balance* in insurance (Denuit et al., 2021); intuitively, we need to collect sufficient premia $\hat{Y}(X_i)$ to cover all claims Y_i. In this case of a single group, calling $\hat{Y}_{\text{af}}$ 'actuarially fair' is some abuse of naming, since in this case it corresponds to full solidarity.

Proposition 4.5. *Given a partition $\mathcal{G} \subset 2^{\mathcal{X}}$ of $\mathcal{X}$, the actuarially fair predictor $\hat{Y}_{\text{af}}$ satisfies (theoretical) calibration with respect to $\mathcal{G}$, and is furthermore the coarsest calibrated predictor in the sense that any other predictor which is calibrated with respect to $\mathcal{G}$ either coincides with it or is not group-wise constant.*

[17] Accordingly, probability refers to indicator gambles, that is, indicator functions of events.

The trivial proof is in Appendix C. If we had access to the true outcomes Y, we could use them for the finest, calibrated predictions. In this way, calibration is still a coarser criterion as it is based on the expectation, the theoretical average, and thus aligned with actuarial fairness; perfect accuracy is in general not demanded. In the extreme, making the partition finer and finer we reach segments of one. Calibration then demands that $\hat{Y}(x) = \mathbb{E}[Y|X = x]$, which we called perfect actuarial fairness. In this way, we obtain a 'spectrum of calibration', where refining the choice of groups interpolates between two extremes. Group-based actuarial fairness attempts to approximate the extreme of segments of one given practical constraints. Hence, actuarial fairness is well-aligned with fairness-unaware machine learning, where the goal is to approximate the conditional expectation $\mathbb{E}[Y|X]$ as closely as possible. In the limit, the distinction between group-based approaches to fairness and *individual fairness* (in the sense of the machine learning literature (Dwork et al., 2012)) then becomes blurry: perfect actuarial fairness corresponds to the notion of *individual merit* (Joseph et al., 2016), but in line with Binns (2020) we argue in Section 5 that this does not yield actually 'individual' fairness.

Actuarial fairness is closely related to calibration due to Proposition 4.5; however finer predictors (e.g. the perfect predictor) satisfy calibration, as well. For a precise conceptual correspondence with perfect actuarial fairness, we could consider the following class of fairness measures, inspired and slightly generalized from Räz (2021).[18]

Definition 4.6. *A fairness measure is probabilistically conservative if it is necessarily satisfied by the perfect actuarially fair predictor* $\hat{Y}_{\text{af}}(x) = \mathbb{E}[Y|X = x]$.

For simplicity and due to the insurantial interpretation, we focus on calibration.

Not only for calibration, but similarly for independence (Definition 4.1) the choice of grouping is crucial. At the one extreme, choosing S to be the indicator of the whole population, independence becomes vacuous as it is trivially satisfied. In contrast, we can add more and more groups (sensitive features) for which we demand independence, so that less and less variations in $\hat{Y}$ are allowed. In the limit, then, we reach *full solidarity* with a constant $\hat{Y}$ (see Appendix C.1). Contrasting independence and calibration, the question is whether we allow the prediction $\hat{Y}$ (the premium) to be sensitive to a certain group or not — where the within-group variation is however neglected.

Calibration and independence can be mapped onto *responsibilization vs. non-responsibilization*. The fairness notion that is embodied by calibration is that of accurately reflecting the 'true' probabilities, that is, holding individuals responsible for their risk. In contrast, independence works to decouple predictions from 'true' probabilities to some extent and thus can be viewed as non-responsibilizing — independence is similar, albeit not equivalent, to affirmative action (Räz, 2021). As a consequence, we expect different *performative effects* (see Appendix D), i.e. downstream effects, when applying calibration vs. independence.

For practice, a simple suggestion is as follows. Since calibration comes practically for free for loss-minimizing predictive models (Barocas et al., 2019, p. 62f), we may focus on demanding certain independence relationships. Assume that we have designated a subset of features X_R which we aim to responsibilize for, and a set of features X_{NR} which we aim *not* to responsibilize for. Hence $X = (X_R, X_{NR}, X_{\text{other}})$. Conditional independence (Castelnovo et al., 2022) then demands that

$$\hat{Y} \perp\!\!\!\perp X_{NR} \mid X_R$$

The features X_{other}, on which we withhold judgement, can then be used by the model in a way restricted by the conditional independence. In the spirit of Section 3.2, however, the choice of features for (non)responsibilization should be the outcome of a reflexive process of inquiry.

Beyond calibration and independence, other proposals have been put forward in the machine learning literature, which may also be linked to (non)responsibilization; hence the lessons from insurance can be applied, too. For instance, within an equality of opportunity framework, Heidari et al. (2019) suggest splitting the whole set of features

[18] Räz (2021) defines a fairness measure as conservative if it is necessarily satisfied by the perfect predictor $\hat{Y} = Y$.

into a set of "accountability" features and "irrelevant" features. From our perspective, this maps onto responsibilization and non-responsibilization. As another prominent example, in a causal fairness framework, Kilbertus et al. (2017) assume that a set of "resolving variables" is given, which are influenced by a sensitive feature in a way that it considered "non-discriminatory"; but the authors do not provide guidance on how to choose them. We believe that the lessons from insurance about causality and control (Section 3.1), and more broadly on responsibility in general, can on the one hand guide the selection of such features. On the other hand, we have seen that causal and control-based dimensions of responsibility are highly sensitive to social context (Section 3.2). This highlights the normative element in making such a distinction (responsibilizing or not), which can be problematic and must be recognized as such. The choice can be side stepped by favoring solidarity over actuarial fairness.

5 Tensions between Aggregate and Individuals

Throughout the history of insurance, and also highly relevant to machine learning, we find tensions between *aggregate and individual*. The mode of the welfare states operates with the imaginary of a collective, in which the individual is mutualized in solidarity. This aggregate viewpoint, where an individual is always identified with the average of some group, finds continuity in group-based actuarial fairness (Thiery & Van Schoubroeck, 2006) — consistent with Quetelet's *average human*. Social movements, however, argued that the group-based actuarially fair price is not fair from the viewpoint of the individiual — the desire was to "navigate the social world *unmarked* by the social stereotypes (fashioned by actuarial science as 'objective' statistical classifications) [..]" (Krippner, 2023, emphasis in original). This critique has prompted a shift in the enactment of actuarial fairness (Meyers & Van Hoyweghen, 2018), giving rise to the individualization and personalization of risk and thereby threatening (risk subsidizing) solidarity by dissolving the aggregate. At the heart of this aggregate vs. individual tension is the multifaceted concept of responsibility: what links the individual to the aggregate is mutualization based on establishing shared responsibility. *Personhood*, being an individual, is deeply intertwined with assigning responsibility, as insurance scholars have pointed out (McFall & Moor, 2018; Moor & Lury, 2018). It is never the whole of a person that a premium is attached to in insurance, but specific, contextually relevant aspects (McFall & Moor, 2018), thereby transforming a person into an insurance risk (Van Hoyweghen, 2014; McFall & Moor, 2018). It would be intriguing to explore how personhood is negotiated within machine learning systems, drawing parallels with similar studies in insurance (e.g. (Tanninen, 2020)).

The mode of machine learning is a paradoxical one: on the one hand, it fits with the mode of individualization and personalization. The goal is to provide highly tailored predictions for the individual. On the other hand, aggregates are central to the workings of machine learning: they appear in the input data due to categorization processes; second, the fairness of machine learning systems is typically evaluated based on groups (with the exception of *individual fairness*, see below); third, machine learning in general, whether fairness-unaware or not, rests on aggregate criteria such as average training error. We suggest that a large share of the social worries and issues surrounding machine learning can be understood by framing them in the context of the aggregate vs. individual tension.

Group-based actuarial fairness, which relies on historical data aggregated by groups, is prone to reproduce past injustice (Daniels, 1990; Lehtonen & Liukko, 2015), see also Appendix D. In contrast, the allure of perfect actuarial fairness associated with the personalization of risk, driven by big data and machine learning, is that it is supposedly *individually fair* — the goal being 'segments of one' and setting the premium as $\mathbb{E}[Y|X = x]$. However, we contend that this elusive goal cannot be reached. The core issue lies with the 'hidden collective'. The working of a neural network is similarity-based computation, arguably interpolation (Hasson et al., 2020). Predictions are invariably grounded in data from individuals *similar to you*, where the similarity is with respect to the opaque nonlinear character of the network. This argument has been made in the context of insurance: 'individualized' risk is still relative to the other members of the collective (Tanninen, 2020; Prainsack & Van Hoyweghen, 2020). Yet individual justice in an Aristotelian tradition requires treating people *as individuals* (Thiery & Van Schoubroeck,

2006; Jorgensen, 2022), not based on the data of others.[19] For the same reason, what is called *individual fairness* in machine learning fails to be genuinely individual, as pointed out by Binns (2020). Problematically the collective is implicit, hidden, in the mode of personalization; without transparency and explainability, individuals cannot recognize their own context. Insurance scholars have also argued that this diminishes opportunity for collective action (Moor & Lury, 2018; McFall & Moor, 2018; Krippner & Hirschman, 2022) — the study of collective action in machine learning has just begun (Hardt et al., 2023).

Another way of framing this consists in problematizing the conceptual foundation of probability and statistics itself. Besides the subjective variant of probability, which we consider unfit for decision making affecting people,[20] probability and statistics are fundamentally based on aggregates (Desrosières, 1998). Currently, no viable concept of individual probability is available (Dawid, 2017); instead, probability relies on a reference class (Reichenbach, 1949; Hájek, 2007). While multi-calibration aims at finer aggregates, it is still not individual (Dawid, 2017). Thus, even speaking of an 'individual probability', which perfect actuarial fairness aims at, has no sound conceptual basis. In fact, Friedman (2020, Chapter 4) provides an insightful account for the close link of frequentist probability and insurance in the solidaristic mode of the welfare state. This account invites us to consider a reference class as a class of solidarity, which implies disregarding the quest for the single 'right' reference class and instead recognizing the normative element in this choice. Furthermore, randomness can then also be viewed as a normative assumption in the face of uncertainty, establishing shared responsibility: "anyone of us could have had the accident". Thus, the kind of data that Venn (1876) has in mind, combining aggregate regularity with local irregularity (randomness), corresponds to the prerequisites for insurance. As a consequence, we find normative character in frequentist probability itself and a link to aggregate-based solidarity. Attempting to individualize frequentist probability then raises a paradox. In the context of insurance, Frezal & Barry (2020) have argued that the actuarially fair expected value is only adequate from the economic, aggregate viewpoint of the insurer, but conceptually inadequate (and hence in particular not necessarily 'fair') for the individual (see also (Frezal, 2016)). Or, in the words of Abraham (1985): "No one has a true expected loss". When taken to the limit, actuarial fairness thus undermines the logic of insurance itself (Heras et al., 2020). In the context of machine learning, we argue that it is problematic to evoke the idea of individual probability (risk), particularly when stakes are high such as in the COMPAS case (Angwin et al., 2016), even when individual probability is beyond reach and has no conceptual foundation to rest on. We thus encourage more modesty about the epistemic potential of machine learning.

6 Related Work

Conceptual literature at the intersection of fair machine learning and insurance is sparse. Donahue & Barocas (2021) study the problem of *externalities of size* by taking inspiration from insurance and challenge the clear distinction between actuarial fairness and solidarity as a consequence. Loi & Christen (2021) study the interplay of machine learning and (non)discrimination in an insurance setting from a normative, philosophical perspective. Frees & Huang (2023) and Charpentier (2022) provide broad overviews on discrimination in insurance and consider implications of using machine learning. Xin & Huang (2023) link formal fairness definitions to insurance on a technical level. The closest work to ours that we are aware of is by Barry & Charpentier (2022), who investigate how the use of machine learning in insurance is related to classical fairness debates, but they set other foci. A general difference from ours to related work is that we do not study the use of machine learning *in* insurance, but are interested in a more abstract conceptual linkage. We also note that Frezal & Barry's (2020) critique of actuarial fairness was a major source of inspiration to us.

[19] We leave open the question to what extent this can be realized by human decision makers.

[20] Setting insurance premia based on subjective probabilities seems objectionable when it affects the welfare of people; similarly for machine learning.

7 Discussion

The main claim of our work is that insurance is an insightful analogon for the social situatedness and impact of machine learning systems. By traversing this conceptual bridge, machine learning scholars can make use of the rich and interdisciplinary literature on insurance. In particular, we suggest that the multifaceted concept of responsibility, tightly linked to causality and control, deserves more attention. We have illustrated problems with actuarial fairness as a notion of fairness in the normative sense. In this way, our suggestions are in line with others who demand moving beyond formal fairness to substantive fairness (Green, 2022) and argue that accurate predictive models need not be 'fair' (Eidelson, 2021).

While in this text we have focused on social issues, there are also technical lessons that machine learning could take from insurance, a technology for handling uncertainty. For instance, the problems of *dataset shift* and *model ambiguity* have been recognized in insurance as well as machine learning; for contributions from insurance see e.g. (Milevsky et al., 2006; Cabantous, 2007; Pichler, 2014; Dietz & Niehörster, 2021). On the other hand, the use of machine learning in insurance is increasing. We thus believe that a research agenda linking machine and learning and insurance may lead to a fruitful, two-way interaction of these fields.

In summary, we offer the following insights. Recent impossibility theorems in the fair machine learning literature (Kleinberg et al., 2017) are not as surprising when considering them in the light of the old, fundamental tension in insurance between solidarity and actuarial fairness. In essence, this tension is grounded in how individuals are related to the aggregates they form. This relation rests on responsibilization. Responsibility and responsibilization should be conceptually distinguished, even if in the recent mode of personalized insurance (tightly linked to machine learning) the two are increasingly intertwined. For insurance and machine learning purposes, reasoning about responsibility crucially requires reasoning about causality and control. We have emphasized that the relevant control question has a normative flavour, and thus cannot be left to engineers alone. What is under individuals' control is often hotly contested. As a general research heuristic, many case studies by social scholars of insurance can inspire analogous studies in the context of machine learning, covering a diverse set of topics; mining the literature is thus a rich source of inspiration.

Acknowledgments

This work was was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich. Thanks to Benedikt Höltgen, Sebastian Zezulka, Renate Baumgartner and Maiju Tanninen for helpful discussions and comments.

References

Kenneth S. Abraham. Efficiency and fairness in insurance risk classification. *Virginia Law Review*, 71(3):403–451, 1985.

Martin Marchman Andersen and Morten Ebbe Juul Nielsen. Luck egalitarianism, universal health care, and non-responsibility-based reasons for responsibilization. *Res Publica*, 21:201–216, 2015.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`. Accessed on June 2, 2023.

Chris Armstrong. Equality, risk and responsibility: Dworkin on the insurance market. *Economy and Society*, 34(3): 451–473, 2005.

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5): 941–973, 1963.

John L. Austin. *How to Do Things With Words*. Harvard University Press, Cambridge, 1962.

Ronen Avraham. Discrimination and insurance. In *The Routledge Handbook of the Ethics of Discrimination*. Routledge, 2018.

Ronen Avraham, Kyle D. Logue, and Daniel Schwarcz. Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195–274, 2013. URL `https://scholarship.law.umn.edu/faculty_articles/576`. Accessed on June 2, 2023.

Tom Baker. On the genealogy of moral hazard. *Texas Law Review*, 75:237, 1996.

Tom Baker. Insuring morality. *Economy and Society*, 29(4):559–577, 2000.

Tom Baker. *Risk, insurance, and the social construction of responsibility*. University of Chicago Press, 2002.

Tom Baker and Jonathan Simon. *Embracing risk: The changing culture of insurance and responsibility*. University of Chicago Press, 2002.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019. URL `http://www.fairmlbook.org`. Accessed on January 5, 2024.

Laurence Barry. The rationality of the digital governmentality. *Journal for Cultural Research*, 23(4):365–380, 2019.

Laurence Barry. Insurance, big data and changing conceptions of fairness. *European Journal of Sociology/Archives Européennes de Sociologie*, 61(2):159–184, 2020.

Laurence Barry and Arthur Charpentier. The fairness of machine learning in insurance: New rags for an old man? *arXiv preprint arXiv:2205.08112*, 2022.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 514–524, 2020.

Joar Björk, Gert Helgesson, and Niklas Juth. Better in theory than in practise? challenges when applying the luck egalitarian ethos in health care policy. *Medicine, Health Care and Philosophy*, 23:735–742, 2020.

Ivan Boldyrev and Ekaterina Svetlova. *Enacting dismal science: New perspectives on the performativity of economics*. Springer, 2016.

Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, 2000.

Nicolas Brisset. Economics is not always performative: some limits for performativity. *Journal of Economic Methodology*, 23(2):160–184, 2016.

Laure Cabantous. Ambiguity aversion in the field of insurance: Insurers' attitude to imprecise and conflicting probability estimates. *Theory and Decision*, 62(3):219–240, 2007.

Michel Callon (ed.). *The laws of the markets*. Oxford: Blackwell, 1998.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12, 2022. Article number: 4209.

Alberto Cevolini and Elena Esposito. From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society*, 7(2):1–11, 2020.

Alberto Cevolini and Elena Esposito. From actuarial to behavioural valuation. the impact of telematics on motor insurance. *Valuation Studies*, 9(1):109–139, 2022.

Arthur Charpentier. Insurance: Discrimination, biases & fairness. *Opinions & Debates*, 2022. URL `https://www.institutlouisbachelier.org/en/insurance-discrimination-biases-fairness/`. Accessed on June 2, 2023.

Karen A. Clifford and Russel P. Iuculano. AIDS and insurance: the rationale for AIDS-related testing. *Harvard Law Review*, 100(7):1806–1825, 1987.

Dan Cooper and Brian Grinder. Probability, gambling and the origins of risk management. *Financial History Magazine*, 93:10–11, 2009.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Norman Daniels. Insurability and the HIV epidemic: ethical issues in underwriting. *The Milbank Quarterly*, 68 (4):497–525, 1990.

Lorraine Daston. *Classical Probability in the Enlightenment*. Princeton University Press, 2023.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

Bruno de Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 1974/2017.

Mitchell Dean. Risk, calculable and incalculable. *Soziale Welt*, pp. 25–42, 1998.

Michel Denuit, Arthur Charpentier, and Julien Trufin. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497, 2021.

Alain Desrosières. *The politics of large numbers: A history of statistical reasoning*. Harvard University Press, 1998.

Rainer Diaz-Bone and Emmanuel Didier. Introduction: The sociology of quantification — perspectives on an emerging field in the social sciences. *Historical Social Research*, 41(2):7–26, 2016.

Simon Dietz and Falk Niehörster. Pricing ambiguity in catastrophe risk insurance. *The Geneva Risk and Insurance Review*, 46(2):112–132, 2021.

Kate Donahue and Solon Barocas. Better together? how externalities of size complicate notions of solidarity and actuarial fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, pp. 185–195, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Benjamin Eidelson. Patterned inequality, compounding injustice, and algorithmic prediction. *American Journal of Law and Equality*, 1:252–276, 2021.

Richard Ericson, Dean Barry, and Aaron Doyle. The moral hazards of neo-liberalism: lessons from the private insurance industry. *Economy and Society*, 29(4):532–558, 2000.

Wendy Nelson Espeland and Michael Sauder. Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1):1–40, 2007.

Wendy Nelson Espeland and Mitchell L. Stevens. A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, 49(3):401–436, 2008.

François Ewald. Die Versicherungs-Gesellschaft. *Kritische Justiz*, 22(4):385–393, 1989.

François Ewald. Norms, discipline, and the law. *Representations*, 30:138–161, 1990.

Francois Ewald. Insurance and risk. In *The Foucault effect: Studies in governmentality*, pp. 197–210. The University of Chicago Press, 1991.

François Ewald. *L'État providence*. Grasset, 1986.

Edward W. Frees and Fei Huang. The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1): 2–24, 2023.

Sylvestre Frezal. Alea and heterogeneity: the tyrannous conflation, 2016. URL `https://www.chaire-pari.fr/wp-content/uploads/2016/09/Alea-and-Heterogeneity_the-Tyrannous-Co` Accessed on June 14, 2023.

Sylvestre Frezal and Laurence Barry. Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167:127–136, 2020.

Rachel Z. Friedman. *Probable Justice: Risk, Insurance, and the Welfare State*. University of Chicago Press, 2020.

Jill Gaulding. Race, sex and genetic discrimination in insurance: What's fair? *Cornell Law Review*, 80:1646, 1994.

Ackerlof George A. The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, and Lorenz Kruger. *The empire of chance: How probability changed science and everyday life*. Cambridge University Press, 1989.

Brian J. Glenn. Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143, 2003a.

Brian J. Glenn. Risk, insurance, and the changing nature of mutual obligation. *Law & Social Inquiry*, 28(1): 295–314, 2003b.

Ben Green. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, 35, 2022. Article number: 90.

Francisca Grommé and Stephan Scheel. Doing statistics, enacting the nation: The performative powers of categories. *Nations and nationalism*, 26(3):576–593, 2020.

Ella Hafermalz, Kai Riemer, and Sebastian Boell. Enactment or performance? a non-dualist reading of Goffman. In *Beyond Interpretivism? New Encounters with Technology and Organization: IFIP WG 8.2 Working Conference on Information Systems and Organizations, IS&O 2016*, pp. 167–181. Springer, Cham, 2016.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156:563–585, 2007.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22969–22981, 2022.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. 2023.

Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 181–190, 2019.

Carol Anne Heimer. *Reactive risk and rational action: Managing moral hazard in insurance contracts*. University of California Press, 1985.

Antonio J. Heras, Pierre-Charles Pradier, and David Teira. What was fair in actuarial fairness? *History of the Human Sciences*, 33(2):91–114, 2020.

Benedikt Höltgen and Robert C. Williamson. On the richness of calibration. In *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency*, pp. 1124–1138. 2023.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398, 2018.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513, 2020.

Robert Huseby. Can luck egalitarianism justify the fact that some are worse off than others? *Journal of Applied Philosophy*, 33(3):259–269, 2016.

Renée Jorgensen. Algorithms and the individual in criminal law. *Canadian Journal of Philosophy*, 52(1):61–77, 2022.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, volume 29. 2016.

Atoosa Kasirzadeh. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 349–356, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems*, volume 30, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pp. 43:1–43:23, 2017.

Carl Knight. Luck egalitarianism. *Philosophy Compass*, 8(10):924–934, 2013.

Greta R. Krippner. Unmasked: A history of the individualization of risk. *Sociological Theory*, pp. 83–104, 2023.

Greta R. Krippner and Daniel Hirschman. The person of the category: the pricing of risk and the politics of classification in insurance and credit. *Theory and Society*, 51(5):685–727, 2022.

Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*, 2021.

Xavier Landes. The normative foundations of (social) insurance: Technology, social practices and political philosophy. 2013. URL `https://www.centroeinaudi.it/images/abook_file/WP-LPF_6_2013_Landes.pdf`. Accessed on June 14, 2023.

Xavier Landes. How fair is actuarial fairness? *Journal of Business Ethics*, 128:519–533, 2015.

Xavier Landes and Nils Holtug. Insurance, equality and the welfare state: Political philosophy and (of) public insurance. *Res Publica*, 21:111–118, 2015.

Sharon M. Lee. Racial classifications in the US census: 1890–1990. *Ethnic and Racial Studies*, 16(1):75–94, 1993.

Turo-Kimmo Lehtonen and Jyri Liukko. The forms and limits of insurance solidarity. *Journal of Business Ethics*, 103:33–44, 2011.

Turo-Kimmo Lehtonen and Jyri Liukko. Producing solidarity, inequality and exclusion through insurance. *Res publica*, 21(2):155–169, 2015.

Turo-Kimmo Lehtonen and Ine Van Hoyweghen. Editorial: Insurance and the economization of uncertainty journal. *Journal of Cultural Economy*, 7(4):532–540, 2014.

Kasper Lippert-Rasmussen. Genetic discrimination and health insurance. *Res Publica*, 21(2):185–199, 2015.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158, 2018.

Jyri Liukko. Genetic discrimination, insurance, and solidarity: an analysis of the argumentation for fair risk classification. *New Genetics and Society*, 29(4):457–475, 2010.

Michele Loi and Markus Christen. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, 34:967–992, 2021.

Deborah Lupton. The diverse domains of quantified selves: self-tracking modes and dataveillance. *Economy and Society*, 45(1):101–122, 2016.

Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. MIT Press, 2008.

Donald MacKenzie, Fabian Muniesa, and Leung-Sea Siu (eds.). *Do economists make markets?: on the performativity of economics*. Princeton University Press, 2008.

Uskali Mäki. Performativity: Saving austin from mackenzie. In *EPSA11 perspectives and foundational problems in philosophy of science*, pp. 443–453, 2013.

Liz McFall. A 'good, average man': Calculation and the limits of statistics in enrolling insurance customers. *The Sociological Review*, 59(4):661–684, 2011.

Liz McFall. Personalizing solidarity? the role of self-tracking in health insurance pricing. *Economy and Society*, 48 (1):52–76, 2019.

Liz McFall and Liz Moor. Who, or what, is insurtech personalizing?: persons, prices and the historical classifications of risk. *Distinktion: journal of social theory*, 19(2):193–213, 2018.

Liz McFall, Gert Meyers, and Ine Van Hoyweghen. Editorial: The personalisation of insurance: Data, behaviour and innovation. *Big Data & Society*, 7(2):1–11, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Andrea Mennicken and Wendy Nelson Espeland. What's new with numbers? sociological approaches to the study of quantification. *Annual Review of Sociology*, 45:223–245, 2019.

Gert Meyers and Ine Van Hoyweghen. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438, 2018.

M. A. Milevsky, S. D. Promislow, and V. R. Young. Killing the law of large numbers: Mortality risk premiums and the sharpe ratio. *Journal of Risk and Insurance*, 73(4):673–686, 2006.

Michael J. Miller. Disparate impact and unfairly discriminatory insurance rates. In *Casualty Actuarial Society E-Forum, Winter 2009*, 2009.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

Annemarie Mol. *The body multiple: Ontology in medical practice*. Duke University Press, 2002.

Liz Moor and Celia Lury. Price and the person: Markets, discrimination, and personhood. *Journal of Cultural Economy*, 11(6):501–513, 2018.

G. Cristina Mora. *Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American*. University of Chicago Press, 2014.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609, 2020.

Alois Pichler. Insurance pricing under ambiguity. *European Actuarial Journal*, 4(2):335–364, 2014.

Barbara Prainsack and Ine Van Hoyweghen. Shifting solidarities: Personalisation in insurance and medicine. *Shifting solidarities: Trends and developments in European societies*, pp. 127–151, 2020.

Jarkko Pyysiäinen, Darren Halpin, and Andrew Guilfoyle. Neoliberal governance and 'responsibilization' of agents: reassessing the mechanisms of responsibility-shift in neoliberal discursive environments. *Distinktion: Journal of Social Theory*, 18(2):215–235, 2017.

Tim Räz. Group fairness: Independence revisited. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, pp. 129–137, 2021.

Lisa Rebert and Ine Van Hoyweghen. The right to underwrite gender: The goods & services directive and the politics of insurance pricing. *Tijdschrift Voor Genderstudies*, 18(4):413–431, 2015.

Florian Rechfeld. Personalised genetic testing and its impact to insurance. *Swiss Re*, 2016. URL `https://www.swissre.com/dam/jcr:24995a5d-5b66-42ea-a2b9-660458bc6e26/Personalised_genetic_test` Accessed on June 14, 2023.

Hans Reichenbach. *The Theory of Probability: An Inquiry Into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, 1949.

Pierre Rosanvallon. *The New Social Question: Rethinking the Welfare State*. Princeton University Press, 2000.

Pola Schwöbel and Peter Remmers. The long arc of fairness: Formalisations and ethical discourse. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, pp. 2179–2188, 2022.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 59–68. Association for Computing Machinery, 2019.

Ronen Shamir. The age of responsibilization: On market-embedded morality. *Economy and Society*, 37(1):1–19, 2008.

Tamar Sharon. Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philosophy & Technology*, 30(1):93–121, 2017.

Deborah A. Stone. Ad missions. The American Prospect, 2001. URL `https://prospect.org/health/ad-missions/`. Accessed on May 17, 2023.

Rick Swedloff. Risk classification's big data (r) evolution. *Connecticut Insurance Law Journal*, 143:339–373, 2014.

Maiju Tanninen. Contested technology: Social scientific perspectives of behaviour-based insurance. *Big Data & Society*, 7(2):1–14, 2020.

Yves Thiery and Caroline Van Schoubroeck. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211, 2006.

Ine Van Hoyweghen. On the politics of calculative devices: performing life insurance markets. *Journal of Cultural Economy*, 7(3):334–352, 2014.

Ine Van Hoyweghen. Genomics and insurance: The lock-in effects of a politics of genetic solidarity. In *Handbook of Genomics, Health and Society*, pp. 203–211. Routledge, 2018.

Ine Van Hoyweghen, Klasien Horstman, and Rita Schepers. Making the normal deviant: The introduction of predictive medicine in life insurance. *Social Science & Medicine*, 63(5):1225–1235, 2006.

Ine Van Hoyweghen, Klasien Horstman, and Rita Schepers. Genetic 'risk carriers' and lifestyle 'risk takers'. which risks deserve our legal protection in insurance? *Health Care Analysis*, 15:179–193, 2007.

John Venn. *The Logic of Chance*. MacMillan, 1876.

Roel Verbelen, Katrien Antonio, and Gerda Claeskens. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304, 2018.

Ed Vosselman. The 'performativity thesis' and its critics: Towards a relational ontology of management accounting. *Accounting and Business Research*, 44(2):181–203, 2014.

Kate Vredenburgh. Fairness. In *The Oxford Handbook of AI Governance*. Oxford University Press, 2022.

Michael A. Walters. Risk classification standards. In *Proceedings of the Casualty Actuarial Society*, volume 68, pp. 1–18, 1981.

David Wilkie. Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1357):1039–1044, 1997.

Jon Williamson. A dynamic interaction between machine learning and the philosophy of science. *Minds and Machines*, 14(4):539–549, 2004.

Xi Xin and Fei Huang. Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, pp. 1–35, 2023.

A Appendix

B Non-Responsibility Based Reasons for Responsibilization

The difference between responsibility and responsibilization can be subtle. In insurance, two classical principles supply reasons for justifying responsibilization *without* being based on responsibility, however. We might also call them "efficiency-based" reasons for responsibilization (Andersen & Nielsen, 2015).

Adverse selection (George A., 1970; Thiery & Van Schoubroeck, 2006; Avraham et al., 2013) refers to the informational asymmetry between insurer and insured (policyholder) at the time of underwriting. Typically, the insured is better informed about their risk; higher-risk individuals are more likely to seek insurance for protection. The reasoning is then that without segmentation, insurers are more likely to attract high-risk individuals who profit from the subsidy of the pool; ultimately, leading to bankruptcy of the insurer. Thus competition drives increasing segmentation. Public insurance has the benefit of compulsory participation, so that adverse selection cannot occur and solidarity can be implemented. In contrast, adverse selection is advanced by the industry as a justification for actuarial fairness (Miller, 2009).

Moral hazard (Heimer, 1985; Baker, 1996; 2000) (in the context of insurance) on the other hand refers to a *performative*, behaviour-shaping aspect of insurance premia. Simply put, the idea is that policyholders are more inclined to behave in a risky way due to the protection offered by insurance coverage. For instance, there is less incentive to purchase precautionary measures such as alarm systems. Like adverse selection, moral hazard is invoked as an argument in favor of actuarial fairness. The welfare state is seen as the ultimate source of moral hazard (Ericson et al., 2000). For instance, public health insurance might lead to more visits to the doctor.[21] Indeed, moral hazard is *the* responsibilization force in the neoliberal era (Ericson et al., 2000). In particular, moral hazard favors behaviour-based personalization (Verbelen et al., 2018). Insurance companies thus increasingly emphasize loss prevention, that is, acting on behaviour (Baker & Simon, 2002; Cevolini & Esposito, 2022). Note that the applicability of moral hazard presupposes control for a feedback loop to exist.

Both adverse selection and moral hazard concern the performative dimensions of "calculative devices" (Van Hoyweghen, 2014). Hence we suggest that these concepts might be usefully transposed onto machine learning.

C Proof of Proposition 4.5

Proof. Plugging in the definition of the actuarially fair predictor, we find that $\forall G \in \mathcal{G}$:

$$\mathbb{E}[\hat{Y}(X)|X \in G] = \mathbb{E}[Y|X \in G] \tag{1}$$

$$\Leftrightarrow \mathbb{E}[(x \mapsto \mathbb{E}[Y|X \in G_x])(X)|X \in G] = \mathbb{E}[Y|X \in G] \tag{2}$$

$$\Leftrightarrow \mathbb{E}[Y|X \in G] = \mathbb{E}[Y|X \in G], \tag{3}$$

[21] Noting the negative connotation in the term *moral hazard*, Baker (1996) asks: "What, after all, is wrong with enabling people to go to the doctor when they feel the need, and why should we be concerned when they do so?"

which is true. For the second statement, observe that Equation 1 implies that a predictor $\hat{Y}$ which is constant on each group G must equal the conditional expectation for that group. □

C.1 Independence with respect to all groups implies full solidarity

Proposition C.1. *Assume independence holds with respect to all groups, that is*

$$\hat{Y} \perp\!\!\!\perp \chi_A, \quad \forall A \subseteq \Omega, A \neq \emptyset.$$

Then $\forall \omega \in \Omega : P(\{\omega\}) > 0 : \hat{Y}(x) = c$ *for some* $c \in \mathbb{R}$.

Proof. Recall that we assume a finite Ω. Then we can assume without loss of generality that $P(\{\omega\}) > 0 \, \forall \omega \in \Omega$, otherwise we could work on an altered probability space by discarding sets of measure zero. From the independence assumption it follows that $P(\{\omega \colon \hat{Y}(\omega) = y | A) = P(\{\omega \colon \hat{Y}(\omega) = y)$ for any $y \in \mathbb{R}$ and $A \subseteq \Omega$, $A \neq \emptyset$. Pick any ω_1 so that $\hat{Y}(\omega_1) = y_1$. Then it must hold $P(\{\omega \colon \hat{Y}(\omega) = y_1 | \{\omega_1\}) = 1 = P(\{\omega \colon \hat{Y}(\omega) = y_1\})$, from which we conclude that $\hat{Y}$ must be constant on Ω. □

D Performativity

A central and unifying theme that emerges when considering social issues surrounding insurance, and as we contend, also machine learning, is that of *performativity*. This concept allows us to comprehend many of the previously raised points through a new lense. The idea of performativity originates from John L. Austin's seminal work "How to do things with words" (Austin, 1962), where Austin notes that the function of language is often not only descriptive, but also has constitutive and causal effects. For example, uttering "I promise" itself *constitutes* a promise and has the causal effect of establishing certain expectations. Austin (1962) refers to sentences with such *performative* force as *speech acts*. Since Austin, the term 'performativity' has travelled far into multiple disciplines and acquired lives of its own, so that it can be hard to pin down precisely a common core (see (Mäki, 2013) for a critique). We will use the term in the broadest possible sense to emphasize similarity of perspectives instead of differences. An influential account has been put forward in economics (Callon, 1998; MacKenzie et al., 2008; MacKenzie, 2008), which has inspired also a recent formalization in machine learning (Perdomo et al., 2020; Hardt et al., 2022). The central claim of this line of work is that economics is not simply in the business of describing or representing an independent, passive reality, but also actively shapes it, for instance by encouraging people to act in accordance with its models (Boldyrev & Svetlova, 2016). Another general framework and a source of inspiration to us, can be found in the work of Mol (2002). To avoid the dualist connotation of the term performance, implying a 'backstage reality',[22] Mol (2002) instead coins the term *enactment*, referring to the multiple and ongoing work that sustains a reality:

> It is possible to say that in practices objects are *enacted*. This suggests that activities take place — but leaves the actors vague. It also suggests that in the act, and only then and there, something *is* — being enacted. [emphasis in original]

Perhaps the simplest way to understand what is at the heart of performativity, we suggest, is to assert that *representation and intervention are entangled* (Vosselman, 2014). Performativity hence contrasts with the commonsense view that perception and action can be neatly separated; in the latter, the task of machine learning is simply to extract patterns from a passive reality 'out there' in an objective way. For instance, Mitchell et al. (2021) distinguish between "world as it is" and "world as it should and could be", mapping onto prediction and decision task. Similarly,

[22] However, for a critical examination of whether this dualism is inherently associated with 'performance', see (Hafermalz et al., 2016).

Kuppler et al. (2021) urge to cleanly separate prediction and decision. A word that frequently occurs in this context is 'bias', which we like to avoid, since we associate it with the notion of an objective 'backstage' reality, whose representation is then distorted.

What then is the relevance of performativity for insurance, and by analogy, machine learning? Actuarial fairness (calibration), or more broadly the fairness of 'accurate' statistical methods, carries with it an aura of objectivity and neutrality. *If* we choose responsibilization, *then* our predictive models better be 'objective and neutral'. Recall that actuarial fairness aims to set premia in accordance with the expected risk for each policyholder, and it is assumed that insurers can know this risk through statistical methods. Glenn (2003a) succinctly captures this as follows:

> [T]here is a general belief that insurance practices are predicated on objective statistics, what has elsewhere been called "the myth of the actuary". The myth of the actuary is the idea that there is a reality in the world that can be captured by rational choice models and statistical analysis—and that insurance companies do this ethically, objectively, and "correctly."

Such objectivity then is supposed to be a source of authority and fairness. In the words of Van Hoyweghen (2014):

> The dominant view is that insurance technologies of risk assessment are somehow 'measuring', 'observing' or 'describing' peoples' insurance risks. This paper calls for a different approach, namely a pragmatist analysis of the *performativity of insurance calculative devices*. Contrary to the financial realism of the everyday categories of insurance numbers, I argue that insurance calculative devices not only represent but generate, intervene and rearrange the worlds in which they are deployed. [emphasis added]

The performativity of insurance and machine learning becomes especially relevant due to ethical implications. Many scholars have argued, providing insightful examples, that insurance is fundamentally a *normative technology* (Baker & Simon, 2002; Glenn, 2003a; Van Hoyweghen et al., 2006; 2007; Lehtonen & Liukko, 2015; Tanninen, 2020; Prainsack & Van Hoyweghen, 2020), depending on causality, control and responsibility. *Doing* insurance or machine learning involves *enacting* certain realities and suppressing others, as we have sketched in Section 3.2. For instance, in the process of collecting data, only some features are considered, and others neglected. Expanding on this, a performativity perspective would emphasize that there is no objective data 'collection' process, that quantification and categorization require significant and ongoing work; such work may be influenced by implicit normative judgements, which becomes ingrained and hidden in the 'representation'. There is now a vibrant, if still nascent, research field on the *sociology of quantification* (including categorization), owing much to the seminal work of Desrosières (1998); for overviews of this field see (Espeland & Stevens, 2008; Diaz-Bone & Didier, 2016; Mennicken & Espeland, 2019), where the reader finds plenty of evidence for such work. Central in this research field is again performativity, or what has been called the *constitutive potential* of quantification (Mennicken & Espeland, 2019). As a noteworthy example, it has been demonstrated that the census, through the introduction of statistical categories, can contribute to the establishment of a collective identity among the individuals it aims to describe (Lee, 1993; Bowker & Star, 2000; Mora, 2014). Thus, a category that was initially intended to merely represent acquires performativity by actively shaping the formation of this particular group (Grommé & Scheel, 2020).[23]

We propose that insurance can act as a model for the performativity of statistical, "calculative devices" (Van Hoyweghen, 2014) that arise from their social situatedness. In machine learning, performativity shows up in at least two ways: on the one hand, it requires training data, and this training data has been shaped by performative forces in a broader social context — referring to the quantification and categorization processes. Using the

[23]To anticipate a criticism, this does of course not imply that arbitrary sets of people can become a mutually recognizing group: the hard conceptual work is to investigate how performativity of groups functions and what its limits are. In Austinian terms, this means delineating the "felicity" conditions, which make performative utterances successful (Brisset, 2016).

data thus imports this performativity into the model. On the other hand, by deploying a machine learning model its predictions may acquire performative force in both a constitutive and causal sense. The predictions may act as interventions and through responsibilization shape the behaviour of people, who for instance may strategically adapt to the predictions — this sense of performativity is also referred to as *reactivity* (Espeland & Sauder, 2007; Espeland & Stevens, 2008), and is closer to the formalization proposed by Perdomo et al. (2020), which emphasizes the causal dimension, but neglects the constitutive one. The extent of this phenomenon, i.e. how much a company can steer a population using a model, has been termed *performative power* (Hardt et al., 2022).

As explicated in Section 4, depending on the choice of fairness metric (or none) and to which features it is applied (i.e. the choice of groups), machine learning can exert responsibilizing and non-responsibilizing force. We suggest that two classes of performative effects can then be broadly distinguished, which is however not a clear dichotomy in light of Section 3.2. When machine learning responsibilizes for a controllable feature, individuals may adapt to the prediction so as to change it — if they receive feedback; this is the hope of personalized insurance, and this setting also motivates the concept of moral hazard (Appendix B). In contrast, blindly applying the principle of actuarial fairness can lead to responsibilizing for non-controllable features, which then runs the risk of reproducing past injustice implicit in the training data. Yet against a background of such past injustice, it is not clear why actuarial *fairness* should be considered as a principle of *justice* — this argument has been made both in the insurance (Daniels, 1990; Lehtonen & Liukko, 2015; Barry, 2020) as well as the machine learning literature (Mitchell et al., 2021; Vredenburgh, 2022; Green, 2022; Kasirzadeh, 2022); see also (Eidelson, 2021) — however, the insurance literature provides illuminating examples. In this way, machine learning (resp. insurance) can implicitly responsibilize for sensitive features such as gender or race; the situation is particularly intricate when a feature is considered as controllable which stands in a constitutive relation to a sensitive feature (Hu & Kohler-Hausmann, 2020), for instance due to performativity.

In response to the performativity of machine learning, we advocate for explicit reflection about how performative forces have shaped the present input data, and furthermore how a model in conjunction with a choice of fairness metric might exert performative force by acting on people. Focusing on (non)responsibilization and performativity implies taking a dynamic perspective. Thus, it becomes imperative to foreground and explicitly model the effects of deploying machine learning systems (Hu & Chen, 2018; Liu et al., 2018; D'Amour et al., 2020; Schwöbel & Remmers, 2022), constrasting with rather static vocabulary such as *bias* or *discrimination*. In this process of reflexive inquiry, we suggest to pay more attention to enactments of causality, control and responsibility — framing them in this way rather than as immutable *facts* implies making them contestable, that is, putting them on the stage for scrutiny (Glenn, 2003a).