CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Modelling conditionally respected social norms: a critique from the intentional stance
Authors	Ross, Don;Wang, Cuizhu
Publication date	2025-07-21
Original Citation	Ross, D. and Wang, C. (2025) 'Modelling conditionally respected social norms: a critique from the intentional stance', Journal of Economic Methodology. https://doi.org/10.1080/1350178X.2025.2535369
Type of publication	Article (peer-reviewed);journal-article
Link to publisher's version	10.1080/1350178x.2025.2535369
Rights	© 2025, the Authors. Published by Informa UK Limited, trading as Taylor & Francis GroupThis is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided theoriginal work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been publishedallow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-10-12 05:56:36
Item downloaded from	https://hdl.handle.net/10468/17749

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Routledge
Taylor & Francis Group

Journal of Economic Methodology

ISSN: 1350-178X (Print) 1469-9427 (Online) Journal homepage: www.tandfonline.com/journals/rjec20

Modelling conditionally respected social norms: a critique from the intentional stance

Don Ross & Cuizhu Wang

To cite this article: Don Ross & Cuizhu Wang (21 Jul 2025): Modelling conditionally respected social norms: a critique from the intentional stance, Journal of Economic Methodology, DOI: 10.1080/1350178X.2025.2535369

To link to this article: https://doi.org/10.1080/1350178X.2025.2535369

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 21 Jul 2025.

Submit your article to this journal

Article views: 12

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rjec20

JOURNAL OF ECONOMIC METHODOLOGY
https://doi.org/10.1080/1350178X.2025.2535369

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Modelling conditionally respected social norms: a critique from the intentional stance

Don Ross[a,b,c] and Cuizhu Wang [d]

[a]School of Society, Politics, and Ethics, University College Cork, Cork, Ireland; [b]School of Economics, University of Cape Town, Cape Town, South Africa; [c]Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta, USA; [d]Faculty of Law and Administration, Jagiellonian University, Kraków, Poland

ABSTRACT

There is a broad consensus in the leading general literature on norms and norm-change that norms are conditional, and based on descriptive and normative expectations. Expectations are a sub-set of beliefs. Hence some primary barriers to norm-change arise from dynamics among beliefs, and between beliefs and preferences. However, the literature has under-examined the distinction between two such barriers, preference falsification and pluralistic ignorance. We clarify the implications of the distinction for two leading conceptual frameworks (due to Kuran and Bicchieri, respectively), and ultimately for explicit models, of norms and norm change. We furthermore explain how, once this clarification is in place, the two models are naturally reconciled by interpreting the concepts of belief – and also preference – that they incorporate as based on attribution from Dennett's intentional stance.

ARTICLE HISTORY
Received 12 January 2025
Accepted 14 July 2025

KEYWORDS
Social norms; conditional expectations; pluralistic ignorance; preference falsification; normative reference networks; the intentional stance

1. Introduction

Though the literature on social norm change involves many areas of conceptual and empirical debate, rough consensus has formed around some foundational assumptions. One is that most people publicly express support for norms, and behaviourally conform to them, *conditionally* on their expectations of similar expressions and support across members of populations in which they live (Bicchieri, 2006, 2017). Another is that the expectations in question involve both *descriptive* and *normative* elements: people choose to support or follow norms based partly on whether they expect that others *will* do likewise, and partly on whether others will manifest *approval* of observed norm endorsement and compliance (Bicchieri, 2006, 2017). Based on this consensus, most scholars agree that norm change in a society depends on social-scale evolution of descriptive and normative expectations, sometimes impeded or accelerated by explicit rules. (We include the restrictions 'in a society' and 'social scale' because an individual or small network might change their governing norms as a result of changing their social locations, without changing any of their social expectations.)

Expectations are a sub-set of beliefs. Consequently, social-scale normative change necessarily involves social-scale doxastic revision, that is, revisions in socially coordinated beliefs or 'public opinion'. Such revision may result from changes in the composition of society, in cases of intergenerational normative change, or it may result from agents (individuals or coordinated sub-networks)

CONTACT Don Ross don.ross931@gmail.com

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

revising their beliefs over time. In either circumstance, change may be blocked or retarded by various doxastic impediments. In this paper, we focus on two such barriers that have received attention: *preference falsification* and *pluralistic ignorance*.[1] Our aim is to distinguish them more clearly than previous literature has done *with respect to* conditional norms, and the patterns of norm change dynamics they impede and influence. The theoretical perspective from which we operate understands beliefs and preferences according to the *intentional stance* of Dennett (1987).

We emphasise that our concern here is exclusively with conceptual engineering to facilitate theoretical modelling; we do not imagine ourselves to be discovering facts about what concepts 'really' mean. The empirical literature on norms is extensive, and like most scientific work, it owes its clarity to careful operationalisation in the context of measured phenomena and effects. Thus the issues we address might appear to scientists as hair-splitting. However, most social science, and particularly economics, still operates with folk concepts of belief and preference according to which they are internal, unobserved ('latent') states of individual people that must be inferred from behaviour or from elicited self-reports.[2] This leads to conundrums around the integration of survey data, in which respondents are asked to report attitudes[3] in isolation from normal social contexts, with analyses of socially embedded expression and action (see, for an example, Funk, 2016). Space for more sophisticated inferences could be opened by closing the lag between folk *internalist* understanding of beliefs and preferences and *externalist* conceptions that have become dominant in cognitive science over the past few decades. In focusing on this difference, we will not be concerned with philosophical issues around the semantics of belief and preference content. Dennett was the first theorist to systematically argue for externalism not on a priori semantic grounds, but on the basis that beliefs and preferences are evolved operational social devices by which people coordinate expectations of others, and, derivately, of themselves.[4] The absence of a crisp distinction between preference falsification and pluralistic ignorance is, we will argue, a diagnostic expression of social scientists' inexplicitness about the internalism / externalism distinction.

We add further legibility to our aim by triangulating it with foundational work by Kuran (1995) and Bicchieri (2006, 2017). Leading empirical work on effects of norms on political change by Valentim (2024a, 2024b) explicitly extends and refines Kuran's model of preference falsification. At the same time, Valentin (2024a, p. 34) explicitly endorses Bicchieri's analysis of norms as conditional on two kinds of expectations. In what follows we draw attention to a tension between Kuran's and Bicchieri's models. As we will show, the Dennettian understanding of beliefs and preferences guides resolution of this tension, thereby offering cleaner conceptual infrastructure for empirical work such as Valentine's, and for future studies that aim to build on it.

2. Preference falsification

The idea of preference falsification in the theory of public opinion derives from Kuran's (1995) model of some important historical cases of normative change, such as the collapse of support for Soviet-style Communist regimes in Eastern Europe in the late 1980s and of the legitimacy of the Shah's government in Iran in 1979. It can also be applied to peaceful but sudden social shifts such as the turn against lunchtime drinking in business cultures near the close of the previous century, or the rapid turn from fringe minority to solid plurality support for same-sex marriage in many countries during the present century.

Kuran argues that *rapidity* of large-scale preference reversal in these cases indicates that many people had been expressing and publicly conforming to norms that they had privately disliked, and which they felt free to openly abandon when it became clear that public falsification of what Kuran calls their 'private' preferences was much more widespread than they had realised. That is, we infer from such rapidity that people had been concealing their true preferences, due to expected social sanctions if these preferences were publicly expressed, but stopped the concealment when they received signals that their prior expectations had become mistaken. On Kuran's account, the pattern of such change in expectations tends to involve gradual and relatively undocumented

leakage of true 'private' preferences by falsifiers, until a tipping point is reached, at which point the old normative structure suddenly and very prominently collapses. Valentim (2024a), applying the model to the unexpected emergence of widespread support for radical or extreme rightwing parties in many countries since 2015, describes political cases in which the exposure mechanism operates more suddenly. Specifically, he argues that politicians who face special cost–benefit ratios from risking sanctions send coordinating signals to falsifiers.

On Kuran's model, agents' public preferences are typically strategic, and reflect tradeoffs among three kinds of utility. *Intrinsic* utility refers to the utility a person would gain from satisfaction of a preference if no one else was influenced by what preference she had. For example, if a country turns suddenly against disposable plastics, investors in alternative packaging enjoy new profits, regardless of whether they approve or disapprove of the social preference shift on normative grounds, and regardless of what opinions they express. *Reputational* utility refers to the net payoff a person receives from others' normative responses to a public preference she manifests, hence the reputation an agent gains from revealing a particular preference in public. *Expressive* utility refers to the gains to a person from expressing her true preference in public, given a meta-preference for sincerity over insincerity. Within a population, there is heterogeneity in the weightings people assign to these arguments in their composite utility functions.

The mechanisms that deliver intrinsic utility do not, by specification of the concept, work through normative expectations. Typically they result from changes in relative prices, of goods or skills, or of shifts in positional rents, that most members of a society cannot influence. Such macro-scale processes can and do influence norms, but these are second-order effects that typically involve substantial lags. Consequently, people have no incentive to make their preference expression strategies conditional on intrinsic utility they would gain or lose from norm changes. Again, this is not an empirical claim, but a stipulation: intrinsic utility is *defined as* whatever changes in distributions of utility occur independently of expectations concerning sanctioning of expression. Kuran consequently sets it aside in his model of normative dynamics.

People who attach unusually high weight to expressive utility (sometimes referred to as 'activists' or 'trend-setters') are responsible for the kind of leakage that often erodes a status quo built on widespread preference falsification.[5] Such conditions are arguably never permanently stable equilibria. As Kuran argues, where erosion does not occur, this is because expression of false preferences sometimes causes people to shift their 'private' preferences to conform with those they avow, thereby eliminating the trade-off between their expressive and reputational utilities. If this pattern describes enough members of a population, the status quo ceases to be based on preference falsification. But where false preferences do not adjust, erosion, as ever more people begin to express their true preferences reduces the reputational benefits from preference falsification, until, at a tipping point, for most people these benefits become costs, and reputational and expressive utility become complements instead of substitutes. The result can be very rapid, deep, social change. O'Toole (2022) provides a thick and compelling historical narrative of an actual such process, which within one generation transformed Irish (Republican) society from subservience to the official values of the Catholic Church into one of the most secular and socially liberal national communities ever observed in the world.

As discussed by Tankard and Paluck (2016), and through thicker case studies in Valentim (2024a, 2024b), elections and political crises can function as information shocks that tip societies out of normative equilibria built on preference falsification. Bursztyn et al. (2020) generate incentivised experimental survey data to support the hypothesis that this dynamic characterised the first US presidential victory of Donald Trump. Crucial to the operation of this mechanism is confidence on the part of voters that their ballots are truly anonymous, allowing them to give publicly visible expression of preferences they would falsify if theses expressions could be linked to their public identities. Trump's victory then sent a signal that nativist, anti-feminist, and anti-diversity opinions were more widely held than had been supposed, so in fact *could* be openly expressed. Such expressions might still attract sanctions in the form of moral disapproval from at least half of the population, but the former falsifiers could see that had unexpected force of numbers, and could

expect protection from the President, other members of his party, and sympathetic mass media such as Fox News.

In general, both the theoretical and empirical literatures have been mutually consistent in framing the concept of preference falsification. This is likely due to the fact that in Kuran (1995) it has a canonical source that provides an explicit model. In terms of Bicchieri's distinction between descriptive and normative expectations, when norms shift as a consequence of preference falsification dynamics, the underlying doxastic dynamics mainly concern normative expectations: beliefs about the reputational effects of public expressions of alternative norms, derived from beliefs about what majorities in communities prefer and will tend to sanction or reward.

3. Pluralistic ignorance

In contrast with preference falsification, there is less clear consensus on the concept of pluralistic ignorance. We will discuss recent vagaries of framing in the next section. Here, we start with the concept's origins. As introduced by Katz et al. (1931), pluralistic ignorance referred to incorrect beliefs by individuals about the distribution of normatively valanced behaviour. In the organisational psychology literature that applied the concept over subsequent decades, the primary mechanism for maintaining pluralistic ignorance was conformity bias (Bissell, 2015), that is, widespread desire by individuals to regard their behaviour, and to have their behaviour regarded by others, as non-deviant. Such conformity bias can cause people to over-ride conflicting motivations *mistakenly* if their beliefs about the actual distribution of preferences in their reference group is false, or if they are in error about the extent to which their reference sample is representative of a larger population with which they normatively self-identify.

A person might, for example, believe that she is having fewer risky social adventures than others, and feel that to be 'normal' she should act against her real level of risk aversion. She might be motivated by concern that she is making fewer friends, or engaging with fewer sexual partners, or sampling fewer stimulants, or enjoying fewer travel experiences, than what she believes to be typical of her demographic reference group. On this conceptualisation, the fundamental doxastic bias concerns descriptive expectations about what others *do*. This bias is then typically entrenched by second-order normative expectations, where others whom the ignorant person consults reflect the same conformity bias back to her. Revision of false descriptive beliefs that constitute pluralistic ignorance tends to mitigate second-order false normative expectations automatically, by implying more accurately representative social reference frames.

Norm change in cases where pluralistic ignorance is a main obstacle tends to be gradual and intermittent, unless attention of most agents can be captured by a common and highly credible corrective source. For example, rates of binge drinking among US College students were reported in 2018 as having gradually declined in correlation with beliefs about its frequency, but with the latter still remaining inaccurate and the former still remaining high relative to the age-adjusted general population (Krieger et al., 2018). In typical cases, individuals change their descriptive expectations, and their behaviour, by encountering and being convinced by accurate social statistics, or because a behavioural option becomes less socially salient and the population 'forgets' the problem. Erosion of pluralistic ignorance by accurate social statistics might have recently become less likely due to the loss of influence of evidence-sensitive journalism to social media, and the tendency of the latter to cluster people in online communities with internally homogenous perceptions (Finnegan, 2023).

An economist should be struck by the fact that many of the 'classic' social phenomena to which the pluralistic ignorance concept was originally applied in the organisational psychology literature concerned cases in which conformity bias was perceived to conflict with risk aversion. To continue with the paradigmatic example from this basic literature, most college-age binge drinkers are aware that the behaviour involves acceptance of more risk than moderate drinking or abstinence. A similar point would apply to social patterns in sexual adventurousness or casual attitudes towards self-discipline in educational or workplace settings. Recognition of this point allows for the possibility that

first-order normative expectations could directly support or undermine pluralistic ignorance to the extent that people directly police one another's risk preferences.

However, to our knowledge this possible mechanism has not in featured social-scientific literature on pluralistic ignorance to date.[6] The existence of common idioms such as 'you only live once' indicates that people do express normative expectations about risk tolerance, but it is not clear that such expectations underlie misperceptions of normative standards about specific behaviours or policies. We conjecture that enforcement of normative risk preferences may be largely confined to institutional pressure from agencies with incentives to control costs stemming from risky behaviour. But empirical research on distributions of individual risk preference show strong stability, both over time and across groups (Harrison & Ross, 2016); so there is no general evidence that such institutional pressure effectively shifts norms, at least with any striking speed.

Reflection on this hypothetical mechanism draws attention to a main difference in application between preference falsification and the 'classic' pluralistic ignorance phenomena. Appeal to preference falsification models is motivated, as in Valentim (2024a), by observation of *rapidity* of social-scale norm change. By contrast, pluralistic ignorance was, at least until recently, more often invoked to explain *lags* in social behavioural response to improvements in social knowledge. Of course, there can be no decisive objection, on empirical grounds, to widening the reference set of pluralistic ignorance phenomena so as to attach more relative weight to reinforcing normative expectations. Perhaps the earlier literature over-anchored on the particular exemplars of student binge drinking and slack commitment to school achievement.[7] In general, we should seek to regiment concepts against drift when and only when such conceptual refinement promotes clearer modelling practice. To this we now turn.

4. Conceptual clarity and explicitness in models of norm change

Our critical objective is neither to criticise nor retrench the concepts of preference falsification and pluralistic ignorance *per se*. Rather, we aim to clarify what can be done with these conceptual tools *within the current philosophical environment that has elevated the centrality to social explanation of norms as conditional expectations.* Kuran's work on preference falsification and the organisational psychology literature on pluralistic ignorance pre-dated this environment. The concepts have since been, quite naturally, co-opted by authors whose central focus is on norms. The unsystematic nature of this co-optation is the intended target of our (constructive) criticism.

In an important review article, social psychologists Tankard and Paluck (2016, p. 191) state that '[p]luralistic ignorance is a phenomenon in which individuals comply with a norm but privately reject it'. The reference to 'private rejection' is closer to preference falsification models than to models of pluralistic ignorance, which depends on confused perception of *normality* but not on explicit rejection of status quo norms. 'Comply' is ambiguous as between behavioural compliance, which is common to both preference falsification and pluralistic ignorance, and expressed normative support, the core of preference falsification. This ambiguity may reflect Tankard and Paluck's endorsement of a concept of 'social norm' that is much looser than the concept analysed by Bicchieri and others, and which harkens back to the framework in which pluralistic ignorance was first conceptualised. Citing Asch (1952) and Sherif (1936), Tankard and Paluck say that 'social norms … are defined as what is typical or desirable in a group or situation' (Tankard & Paluck, 2016, p. 184).

In Bicchieri's (2017) influential recent treatment of practical mechanisms for norm change, she argues (pp. 42–46) for the inadequacy of a view, attributed to Berkowitz and Perkins (1987) and Berkowitz (2005), according to which pluralistic ignorance is best corrected by presenting the ignorant with better evidence about descriptive expectations. Pluralistic ignorance, she insists, also involves mistaken normative expectations. She *also* argues that change efforts depend on the ignorant 'privately' rejecting the norms they falsely believe to be prevalent, implying that they must have been engaging in preference falsification. However, she does not use that concept by name, or cite Kuran in her book.

Bicchieri's general model analytically *identifies* norms with alignment between descriptive and normative expectations. In cases of preference falsification such alignment holds, until the tipping point is approached, and the revealed disalignment is precisely why the normative status quo *suddenly* evaporates. In the classic cases of pluralistic ignorance, however, the two sets of expectations may often fail to align in the first place. In the paradigmatic example of students believing that most others participate more frequently and heavily in binge drinking than statistical evidence bears out, few ever expect to encounter public ratification of the opinion that regular binges are a good thing. Many likely suspect that most members of their reference group behave inconsistently with their expressed normative opinions. Thus false descriptive expectations are not necessarily undermined by normative expectations that conflict with them, and this is part of the reason why information interventions in these kinds of cases typically generate sluggish change at best, rather than sudden flips at tipping points. Which kind of change to expect clearly matters to policy interventions designed either to defend favourably regarded norms or undermine disfavoured ones. Therefore, the distinction should be treated carefully by economists interested in social mechanism design.

We can achieve reflective equilibrium over these cases by organising the concepts such that preference falsification is a feature that sometimes does and sometimes does not support pluralistic ignorance. But given primary interest in norm *change*, then we must attend to the point that, as in Valentim's (2024a) application, tendency to sudden reversal is the diagnostic feature of a norm supported by preference falsification. A unified model of conditional norms and norm change should generate this distinction endogenously.

By a 'unified model', we mean an account that incorporates Kuran's distinction between elements of composite utility, which can sometimes be complements and sometimes be substitutes, along with Bicchieri's distinction between descriptive and normative expectations. A useful demarcation between pluralistic ignorance with and without preference falsification should then be constructed in terms of these elements. The key to this, we will argue, is enhanced sophistication about the cognitive and social nature of expectations, that is, of (forward-looking) beliefs. The sophisticated interpretation to which we refer is that of Dennett (1978, 1987 and elsewhere).

5. Beliefs and preferences

Economists and other social and behavioural scientists often treat beliefs and preferences as internal, private, states of agents that must be inferred by others on the basis of behaviour or sincere declarations. In cognitive science this has generated a substantial literature on the alleged importance of 'mindreading', that is, inference of others' hidden mental states based on their overt behaviour, to social coordination (Nichols & Stich, 2003). In game theory as developed by economists, internalism is reflected in recurrent assumptions about high information costs associated with signalling and screening. When theorists assume that beliefs and preferences are internal states of individuals, unless they are haunted by residual mind–body dualism, they are likely to suppose that beliefs either reduce to or emerge relatively directly from states that could be characterised by neuroscientists (Camerer et al., 2005).

Dennett fundamentally challenged this way of understanding beliefs and preferences, and over the decades since he first did so, the alternative conception he developed has arguably become the dominant one in the philosophy of cognitive science.[8] According to Dennett and his many followers, beliefs and preferences are *attributions* to agents of complex dispositions to respond to environmental contingencies, including, in the case of human agents, verbal probes from other people, in ways that foster appreciation of motivations. Such attributions are crucially sensitive to information about the target's past exposure to experience, but in being culturally evolved to facilitate coordinated understanding – and often coordinated strategies – they have a forward-looking aspect, which to some degree idealises the target's consistency. They thus necessarily transcend the information that even the most powerful achievable brain probe could discover. Human organisms become people – are socialised – by learning to attribute beliefs and preferences to themselves

as if they were external observers of themselves, that is, by adopting the same *intentional stance* (IS) to themselves that they do to others. Tellingly, developmental evidence shows that children formulate intentional narratives about others before they apply such formulations to themselves (McGeer, 2001, 2021).

Dennett's account of the epistemic and ontological status of preferences can readily be brought within a framework familiar to economists. It is effectively a qualitative expression of revealed preference theory, according to which utility functions are constructed as models of observed choices rather than causes of them (Binmore, 2009; Ross, 2014). They must be jointly estimated with subjective beliefs about risky outcome probabilities following Savage (1954) (Andersen et al., 2008) so the Dennettian account of beliefs is also naturally incorporated into this 'quantitative IS'; for technical details see Harrison and Ross (2023) and Alekseev et al. (2024).

Zawidzki (2013) pushes the significance of Dennettian theory for social science a step further, by mustering evidence that mindreading is unable to perform effective coordinating functions outside of two special kinds of context: interactions between individuals with rich and deeply shared personal histories, and contexts in which interpretations are unusually tightly constrained by unambiguous rules. In general, he argues, people use the IS to coordinate by reciprocally *shaping* one another's minds through strategic exertion of influence, rather than by *reading* them and conditioning their own choices on these inferences.

If such *mindshaping* is indeed strategic, then game theorists should aspire to model it. This is challenging, because standard game theory infers vectors of equilibrium strategies from exogenously fixed utility functions of players, whereas mindshaping characterises the main kind of process by which utility functions are formed, and by which, therefore, game structures are determined (Ross, 2005). The history of game theory, however, provides a clue to the way around this impasse through the development of models of Bayesian learning of correlated equilibrium (CE) (Aumann, 1974, 1987; Gintis, 2009). Ross and Stirling (2025) show that strategic play conditioning on players' actions rather than on outcomes can generate CE in subsequently determined games among players whose utility functions are stabilised by the prior process. This achieves a key step in achieving the agenda motivated in Ross and Stirling (2021) to build a formal theory of mindshaping; for integration of this theory with accounts of norm dynamics, see Ross et al. (2024).

What implications does this view of preferences and beliefs (including expectations) suggest for a unified theory of conditional norms? The key general insight is that neither of the main elements of microfoundational accounts of social norm change – beliefs and preferences – have content that can be identified independently of social expectations. People attribute belief and preference content, both to others and to themselves, *for the sake of* rationalising their patterns of socially observable behaviour.

When Kuran (1995), and, following him, Valentim (2024a), refer to 'private' preferences, we should understand 'private' as a modelling idealisation.[9] In applications, we are not asked to imagine normative opinions that are kept hidden from all second-party observers. But the invocation of privacy does invite us to imagine preferences that are relatively hidden from society at large. We will argue that this encourages the equivocations identified above between preference falsification and pluralistic ignorance. Because our topic is conceptual engineering, we will initially follow Kuran's example and focus on idealised purely private beliefs and preferences, then extend the derived insights to more realistic cases.

Dennettian externalism does not impugn the idea of purely private beliefs, because, as discussed above, all socially competent people learn as children to take the intentional stance towards themselves. An ecologically meaningful *network* of beliefs could not exist without immersion in a social context that made most of its elements socially significant, but various *specific* beliefs can be self-attributed secretly. However, purely private opinions cannot be understood as preferences attributed to help coordinate socially relevant choices.

It would exaggerate the point to claim that we can never talk meaningfully about purely private preferences. Most people, for example, probably have a 'preferred' direction of wiping themselves

after defaecation that no one else knows about. But there is no evident social scientific payoff from assimilating such habits, just because ordinary language is tolerant of multiple, loose verbal ontologies, to the phenomena social scientists label 'preferences'. Economists' practice, as noted above, follows IS-theory here: revealed preference models require observable choices (which might be choices about what opinions to express) in the evidence base for utility assignments. So purely private opinions cannot be understood as expectations under IS theory or standard economic analysis.

To preserve a role for any purely private opinions in an account of norms, we would have to interpret expressive utility as deriving entirely from relief of hidden discomfort from insincerity, and not, at least in part, from relief from feelings of alienation from what is perceived to be public opinion. Such thin expressive utility would be unlikely to erode preference falsification until and unless the issues on which preferences were falsified faded in social significance. Expressive utility puts pressure on preference falsification because falsifiers suffer utility loss from feeling estranged from wider communities in which they regard themselves as entitled to membership.

Again, literally private preference should simply be understood as an idealisation in preference falsification models. It is evident from everyday observation that people often conceal the beliefs and preferences that they and their intimates ascribe to themselves from *most* others – that is, they express them in settings where access is restricted to intimates or people to share their minority views, but not in fully public ones. Valentim (2024a) discusses the airing of extreme and radical right-wing political stances in restricted internet environments, where people cultivate and coordinate on right-wing attitude sets, prior to being encouraged by right-wing candidates and parties to stop falsifying their preference expressions in wider public settings.

The relevant distinction, then, between relatively private and relatively public opinions, asks us to attend to norms that operate variably for a given individual across alternative reference networks.

Here we can find the basis for a principled general distinction between preference falsification and (mere) pluralistic ignorance. Preference falsifiers actively express contradictory preferences as their reference networks shift. The resulting *social* (as opposed to merely) cognitive dissonance contributes to the valuation they attach to expressive utility. The point is not simply that all of this preference content is observable. Understanding it using the intentional stance emphasises that its reality largely *consists in* its social explicitness. To extend the example studied by Valentin (2024a), participants in right-wing virtual chat sites likely did not come to form *any* identifiable preferences about their signature themes (e.g. immigration, freedom to express negative stereotypes about ethnic and sexual minorities and women, impatience with legal and bureaucratic policy-making) until their 'hidden' reference networks provided them with an environment in which such preferences could be attributed and self-attributed – and, crucially, *refined* into explicit opposition to more widely prevalent norms. Special normative reference networks are effectively mindshaping hothouses, which *create* conflict between expressive and reputational utility of participants. Though digital social media have no doubt made this phenomenon much more widespread, with highly visible consequences for political and social stability in many countries, the dynamics are not new. Revolutionary nationalist movements in the nineteenth century, underground anarchist, socialist and communist parties in the era of classic industrialising economies, associations of feminist and gay-rights advocates in the twentieth century, were all manifestations of the basic model. Some of these cases were morally progressive and some were not; some (many nationalists, first-wave feminists) generated change in prevailing social norms while others (anarchists, most communists) failed to do so.

Let us now contrast this with mere pluralistic ignorance. College students who think that most other college students enjoy excessive drinking generally are not called upon by their social environments to forge *any* explicit preferences to the effect that all *should* follow this behavioural pattern. Some drink a lot because they enjoy it, while (more) others follow the crowd and drink a lot and find themselves regretting the consequences. From the intentional stance, there is no reason to attribute *normative* preferences to them (about this subject) at all. Later in life, when their reference group has

changed, many attribute reckless preferences to their former selves retrospectively. If their drinking behaviour does not change in adulthood, most express preferences to achieve sobriety *both* within and outside of networks of other substance abusers (Heyman, 2009); after leaving college, the conditions fostering pluralistic ignorance conditions cease to hold. Again, we can readily consider other examples. Rural Chinese women who regarded divorce as rare and exotic move to urban environments and discover that it is much more common than they supposed, thus leading them to consider dissolving their unsatisfactory unions (He, 2021). Men who think that almost no other men trade off paid working hours for childcare responsibility, encounter, if they are lucky in their environments, other fathers in the playground and become less resistant to supporting their partners' career ambitions. In social studies and models of these phenomena, preference falsification is not a significant aspect of dynamics (which is, of course, not to claim that it is non-existent).

We certainly do not mean to suggest that beliefs and preferences are attributable from the intentional stance only when they are expressed. Economists who apply revealed preference theory, the most explicit use of the intentional stance in social science (Harrison & Ross, 2023; Ross, 2014, 2025), typically identify beliefs and preferences on the basis of 'mute' choice behaviour. But the motivation for such attribution is economists' disciplinary focus on *agency* (hence the restriction to *chosen,* meaning incentive-sensitive, behaviour). Consequently, it implicitly appeals to a point recurrently emphasised by Dennett, that belief and preference attributions are grounded in actual or hypothetical *probes* that encourage people to generate coherent narratives about their agency. Probing the *normative* expectations of people caught up in mere pluralistic ignorance will often expose (though perhaps with low behavioural efficacy) misalignment with *descriptive* expectations about norms. This is roughly the opposite effect of probes in cases of preference falsification; preference falsifiers are, more or less by definition, people who are experienced subjects of such probes, and unless they believe that the administrator of the probe is part of their minority reference group, they will typically simply report the false preference.

Dennett (1991) emphasised that the self-narratives that constitute people's consciously constructed selves typically achieve a minimal socially required standard of consistency by tolerating gaps in areas that, if attended to, would signal internal conflict. Such gaps, we suggest, are characteristic of people in classic instances of mere pluralistic ignorance. There need be no fact of the matter as to whether the typical college binge drinker thinks that she or others *should* adopt, or even consider, general normative pro-attitudes towards such drinking. Thus her ignorance goes deeper that unawareness of accurate social statistics; she is ignorant about an entire dimension of potential self-assessment. We conjecture that this phenomenon is *at least as common* as preference falsification, in which people are aware that they explicitly affirm contradictory preferences in different social reference contexts, and may consequently experience disutility from feeling that in more widely exposed environments they are buckling under pressure against manifesting their 'true' selves.

We close by noting that understanding the ontology of beliefs and preferences following Dennett also has implications for Bicchieri's model. Bicchieri (2006) distinguishes 'personal' or 'moral' norms, which people are taken to hold unconditionally of expectations about others, from social norms that are necessarily conditional (see also Valentim, 2024a, p. 36, note 2). However, it is far from clear, either empirically or conceptually, that norms of the first kind are sufficiently common to be interesting in most contexts studied by social scientists. A person who announced adherence to what they called a moral norm that was endorsed by no one else would likely be thought to thereby demonstrate misunderstanding of what morality is. Far from being an interpretable move in a conditional game, such aberrant moral framing would typically carry risk of exclusion from such games. Precisely for this reason, any hypothetical examples we might imagine appear *silly* – consider someone declaring that they consider it morally wrong to sign documents using green ink, or to hop on one foot. Such examples resist clear interpretation from the intentional stance, and are for that very reason almost never observed. A non-psychotic person would struggle to sincerely attribute such a belief *to themselves*.

If Bicchieri's distinction is taken seriously, and held to apply to *actual* norms that many people strongly moralise such as norms against violence, exploitative deception, or standards of civic decency, the consequence would seem to be a severe restriction of the scope of her theory, and of its relevance to many important policy issues. Clearly such restriction is not intended by Bicchieri. One of the leading examples discussed in Bicchieri (2017) is policy intervention to change norms prevailing in several northern African countries that favour genital mutilation of young girls. Beliefs about this practice, both for and against, are typically heavily moralised. We can straightforwardly reinterpret such preferences in exactly the way we reinterpreted 'relatively private' preferences, as preferences which are inconsistently expressed across different reference networks. Then, for example, a Gambian mother who modernises her normative stance with respect to what should happen to her daughter can be understood as shifting her preference from a parochial to a more universal reference network, and if she furthermore comes to express this preference in a traditional social setting, as acting to undermine a dynamic of preference falsification.

Attaching greater emphasis than the leading literature on the general concept of a norm has done to pluralities of normative reference networks in many people's awareness thus supports unification of Kuran's and Bicchieri's conceptual frameworks. The problem of potential misalignment between descriptive and normative expectations in core cases of pluralistic ignorance results simply from unrepresentative sampling, which in turn leads to lack of social pressure against the misalignment. By contrast, preference falsification is straightforwardly understood as arising from conflict between reference networks within a heterogeneous population. It produces sudden and surprising social normative change when people underestimate the size of the reference group in which preference falsification is not encouraged, and this underestimation is exposed by events, activists, or political entrepreneurs.

5. Conclusion

We have used the distinction between descriptive and normative expectations to frame the difference between preference falsification and pluralistic ignorance. This appeals to Bicchieri's general model of norms. We have also endorsed the value of Kuran's three-factor model of utility for understanding the most socially significant examples of norm change (which includes those Bicchieri mislabels as mere pluralistic ignorance cases). This invites a project to combine Bicchieri's and Kuran's explicit models. Such reconciliation need not be motivated merely by an inclination to preserve as many existing theories as we can squeeze into one compartment. The concepts of belief and preference featuring in both models are under-theorised by their authors, and are not explicitly refined to fit the larger economic and game-theoretic research programmes in which they are intended to be embedded. By understanding beliefs and preferences as social constructions from the intentional stance, and understanding norm expression, behavioural adherence, and rebellion as mindshaping moves in conditional games, we achieve reconciliation of the models of norm change while requiring little additional theoretical innovation.

Notes

1. We do not provide encapsulated statements of these concepts at the outset precisely because we do not find that the existing literature has converged on any such statements. Consequently, we could not provide them at the outset without begging our main question; but discussion of the concepts will be extensive in what follows, as they constitute our primary topic.
2. This point applies much more straightforwardly to economics and political science than to sociology or anthropology. But transmission of greater awareness of it among sociologists to other social scientists has been hampered by sociologists' recent suspicion of *any* large-scale theoretical generalisations (see Ross, 2023 for details).
3. The word 'attitudes' denotes different concepts across relevant literatures. Philosophers often use it as an umbrella term for cognitive relationships to propositions, so that beliefs and preferences are species of attitudes. Social psychologists use it to denote generalised evaluative stances to ideas, policies, groups, people, and

cultural practices. To avoid a distracting digression, we will avoid claims that would require us to commit to one usage or the other. We thank an anonymous reviewer for alerting us to this potential source of ambiguity.

4. Of course Dennett had less systematic precursors in developing this idea. See Ross (2025) for relevant history.
5. An anonymous reviewer points out that people who are called 'activists' in everyday language are typically supposed to gain utility directly from their policy campaigns, if these campaigns are successful. This is not expressive utility. Nor, if the activist's own efforts meaningfully change the probability of the policy change, does it fit Kuran's definition of intrinsic utility. In the context of the model, however, 'activist' and 'trend-setter' should simply be taken as heuristic labels for people who value sincere preference expression more than they disvalue expected social sanctions. And by the assumption of the model, an individual activist does not significantly influence the probability of either policy change *or* norm change.
6. There is of course a very large literature on the explicit and official *regulation* of risk, which is obviously normative. There is also a substantial literature in economics measuring risk preferences in various populations. What is missing is work to connect these literatures with social-scale risk-related norms. One plausible retardant to such connection is a widespread view that normative policy should be rationalisable under expected utility theory (Adler, 2012, Chapter 4; but see Buchak 2013), whereas risk-related norms that might spread through social influence are likely to reflect the more widespread actual pattern of risk preferences that violate EUT through subjective rank-dependence (Harrison & Ross, 2016).
7. An anonymous reviewer drew explicit attention to this possible issue in assessing more recent literature by reference to ancestral work.
8. A few highlights in the literature that extends applications of the intentional stance are Clark (1997), Bogdan (1997), Ross (2005), and Mölder (2010).
9. Thanks to an anonymous reviewer for urging us to make this point explicit.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Travel to conferences at which earlier versions of this work was presented was provided by the School of Economics, University of Cape Town (Ross) and the National Science Centre, Poland [grant number 2020/37/B/HS5/00272] (Wang).

ORCID

Cuizhu Wang http://orcid.org/0000-0002-4703-3323

References

Adler, M. (2012). *Well-being and fair distribution*. Oxford University Press.

Alekseev, A., Harrison, G., Lau, M., & Ross, D. (2024). *Deciphering the noise: real patterns in welfare from incentivised choice*. CEAR Working Paper 2024-03. https://cear.gsu.edu/wp-2024-03-deciphering-the-noise-real-patterns-in-welfare-from-incentivized-choice/. Forthcoming in T. Millhouse, S. Petersen, & D. Ross, eds., *Real patterns in science and nature*, MIT Press.

Andersen, S., Harrison, G., Lau, M., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618. https://doi.org/10.1111/j.1468-0262.2008.00848.x

Asch, S. (1952). Group forces in the modification and distortion of judgments. In S. Asch (Ed.), *Social psychology* (pp. 450–501). Prentice Hall.

Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, *1*(1), 67–96. https://doi.org/10.1016/0304-4068(74)90037-8

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, *55*(1), 1–18. https://doi.org/10.2307/1911154

Berkowitz, A. (2005). An overview of the social norms approach. In A. Singhal, M. Cody, E. Rogers, & M. Sabido (Eds.), *Changing the culture of college drinking: A socially situated health communication campaign* (pp. 193–214). Hampton Press.

Berkowitz, A., & Perkins, H. (1987). Current issues in effective alcohol education programming. In J. Sherwood (Ed.), *Alcohol policies and practices on college and university campuses* (pp. 69–85). National Association of Student Personnel Administrators Monograph Series v. 7.

Bicchieri, C. (2006). *The grammar of society*. Cambridge University Press.

Bicchieri, C. (2017). *Norms in the wild*. Oxford University Press.

Binmore, K. (2009). *Rational decisions*. Princeton University Press.
Bissell, J. (2015). Conformity bias and catastrophic social change. In J. Bissell, C. Caiado, S. Curtis, M. Goldstein, & B. Straughan (Eds.), *Tipping points: Modelling social problems and health* (pp. 168–182). Wiley.
Bogdan, R. (1997). *Interpreting minds*. MIT Press.
Buchak, L. (2013). *Risk and rationality*. Oxford University Press.
Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, *110*(11), 3522–3548. https://doi.org/10.1257/aer.20171175
Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, *43*(1), 9–64. https://doi.org/10.1257/0022051053737843
Clark, A. (1997). *Being there*. MIT Press.
Dennett, D. (1978). *Brainstorms*. Bradford.
Dennett, D. (1987). *The intentional stance*. MIT Press.
Dennett, D. (1991). *Consciousness explained*. Little Brown.
Finnegan, C. (2023). *Coordinating minds: mindshaping, communication and technology* [PhD thesis]. University College Cork. https://hdl.handle.net/10468/15936
Funk, P. (2016). How accurate are surveyed preferences for public policies? Evidence from a unique institutional setup. *Review of Economics and Statistics*, *98*(3), 442–454. https://doi.org/10.1162/REST_a_00585
Gintis, H. (2009). *The bounds of reason*. Princeton University Press.
Harrison, G., & Ross, D. (2016). The psychology of human risk preferences and vulnerability to scare-mongers: Experimental economic tools for hypothesis formulation and testing. *Journal of Cognition and Culture*, *16*(5), 383–414. https://doi.org/10.1163/15685373-12342185
Harrison, G. W., & Ross, D. (2023). Behavioral welfare economics and the quantitative intentional stance. In G. W. Harrison, & D. Ross (Eds.), *Models of risk preferences: Descriptive and normative challenges* (pp. 7–67). Emerald.
He, X. (2021). *Divorce in China*. New York University Press.
Heyman, G. (2009). *Addiction: A disorder of choice*. Harvard University Press.
Katz, D., Allport, F., & Jenness, M. (1931). *Students' attitudes; a report of the Syracuse University reaction study*. Craftsman Press.
Krieger, H., Young, C., Anthenien, A., & Neighbors, C. (2018). The epidemiology of binge drinking among college-age individuals in the United States. *Alcohol Research*, *39*(1), 23–30.
Kuran, T. (1995). *Private truths, public lies*. Harvard University Press.
McGeer, V. (2001). Psycho-practice, psycho-theory, and the contrastive case of autism: How processes of mind become second nature. *Journal of Consciousness Studies*, *8*(5–7), 109–132.
McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, *199*(1-2), 1039–1063. https://doi.org/10.1007/s11229-020-02760-7
Mölder, B. (2010). *Mind ascribed*. John Bejamins.
Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford University Press.
O'Toole, F. (2022). *We don't know ourselves*. Norton.
Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. MIT Press.
Ross, D. (2014). *Philosophy of economics*. Palgrave Macmillan.
Ross, D. (2023). Economics is converging with sociology but not with psychology. *Journal of Economic Methodology*, *30*(2), 135–156. https://doi.org/10.1080/1350178X.2022.2049854
Ross, D. (2025). *The intentional stance and revealed preference*. Under review.
Ross, D., & Stirling, W. (2021). Economics, social neuroscience, and mindshaping. In J. Herbecke & C. Herrmann-Pillath (Eds.), *Social Neuroeconomics, Routledge* (pp. 174–201).
Ross, D., & Stirling, W. (2025). Mindshaping, conditional games, and the Harsanyi doctrine. *Journal of Economic Methodology*, *Forthcoming*. https://doi.org/10.1080/1350178X.2024.2413572
Ross, D., Stirling, W., & Tummolini, L. (2024). Modeling norm-governed communities with conditional games: Sociological game-determination and economic equilibria. *Oeconomia*, *14*, 349–398. https://doi.org/10.4000/120ij
Savage, L. (1954). *The foundations of statistics*. Dover.
Sherif, M. (1936). *The psychology of social norms*. Harper Collins.
Tankard, M., & Paluck, E. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022
Valentim, V. (2024a). *The normalization of the radical right*. Oxford University Press.
Valentim, V. (2024b). Political stigma and preference falsification: Theory and observational evidence. *Journal of Politics*, *86*. https://doi.org/10.1086/729966
Zawidzki, T. (2013). *Mindshaping*. MIT Press.