Erasmus Journal for Philosophy and Economics,
Volume 19, Issue 1,
Summer 2026, pp. aa–aa.
https://doi.org/10.23941/ejpe.v19i1.1044

Reconsideration of Bicchieri's Theory of Social Norms: Conceptual Clarifications for Empirical Modelling

CUIZHU WANG
Trinity College Dublin

Abstract: Cristina Bicchieri's theory of social norms is one of the most influential frameworks for diagnosing and measuring norms in behavioural science. Yet its application and experimental practice has revealed conceptual ambiguities and inconsistent operationalisation. This paper offers conceptual clarifications aiming at methodological extensions of Bicchieri's account while remaining largely compatible with the framework. I argue that three aspects of her theory require clarification for effective use in economics: (i) the co-existence of four distinct kinds of social beliefs (first- and second-order, descriptive and normative), (ii) the notion of belief consistency, which should be understood in statistical terms, and (iii) the alignment of reference networks as a condition for norm identification. Refining these aspects provides both improved conceptual clarification and practical guidance for experimental operationalisation. By bridging philosophical analysis with methodological considerations, the paper strengthens the foundations for using Bicchieri's framework to identify and model social norms in economics and the social sciences.

Keywords: Social Norms, Cristina Bicchieri, Social Expectation, Belief Consistency, Reference Network, Norm Modelling

JEL Classification: B41, D83, D91, Z13

I. INTRODUCTION AND MOTIVATION

Cristina Bicchieri's (2006, 2017) theory of social norms is currently the most widely cited and influential framework in contemporary social science for diagnosing and measuring norms. It unites philosophical inquiry and operational specification (Guala 2007, 2017; Ross, Stirling, and Tummolini 2021) and has inspired a large body of empirical research across

AUTHOR'S NOTE: The author notes that this paper was developed from chapter 2 of her PhD thesis, submitted to and successfully defended at University College Cork in 2022. The thesis project was supervised by Professor Don Ross and Professor Glenn Harrison, whose supervision she gratefully acknowledges. The present paper remains the author's sole work; any remaining errors are her own.

economics, psychology, and related fields. Its central appeal lies in offering explicit conditions under which social norms can be identified empirically. Yet despite its prominence, the application of Bicchieri's model in economics and experimental practice reveals persistent conceptual ambiguities and inconsistent operationalisation. These problems risk undermining both its philosophical interpretation and applications in empirical practice.

Three kinds of issues stand out. First, Bicchieri's framework has prompted debate about its psychological commitments, and some of its key constructs admit more than one operational reading in empirical work. In particular, her use of mentalistic concepts has generated debate among philosophers who question whether her model requires too much 'folk psychological machinery' to explain collective normative behaviour (Andrews 2020; Guala 2007; Brennan et al. 2013). These debates matter here insofar as they influence how researchers operationalise the framework's belief condition in experimental design and measurement.[1]

Second, there are tensions between her conceptual analysis and the way it has been operationalised. This is evident both in the types of social beliefs elicited across studies (Huber et al. 2023; Khalmetski 2016; Pérez León 2022; Vriens, Tummolini, and Andrighetto 2023) and in the way belief consistency is measured. For instance, many experimental applications—including Bicchieri and co-authors' own studies (for example, Bicchieri and Xiao 2009; Bicchieri and Chavez 2010, 2013; Bicchieri, Xiao, and Muldoon 2011; Bicchieri and Zhang 2012)—have elicited modal beliefs rather than distributions, effectively equating convergence on point estimates with 'belief consistency'. This methodological shortcut overlooks that consistency is best characterised by distributional features (such as concentration, dispersion, and potential polarisation) of belief responses within a reference network, which are not captured by reporting modes alone (Peterburs et al. 2017; House and Tomasello 2018).

Third, reference networks are often underspecified in empirical designs, even though they are central to Bicchieri's diagnosis conditions; this can matter when populations are heterogeneous or when 'others' is interpreted differently across participants. For instance, in field studies

[1] I have argued elsewhere (Wang 2022, 2023) that aligning Dennett's (1987) theory of the intentional stance (IS) with revealed preference theory (RPT) offers a more promising framework for applying Bicchieri's account to empirical norm modelling. For further discussion of the alignment between IS and RPT, see also Ross (2005, 2014, 2022). In this paper, however, I set that task aside and instead adopt a descriptive approach in reconsidering Bicchieri's model.

of health practices (such as breastfeeding), what participants take 'others' to mean may track a specific local reference network (for example, mothers-in-law and older women) as a subset of, rather than the entire 'community' or a national population—so norm diagnosis can change once the relevant reference network is specified, and that reference networks should not be to treated as coextensive to social groups.[2] Bicchieri (2017) warns that without careful mapping of reference networks, normative interventions risk failure. Yet in much literature, this element is either weakly specified or ignored, leaving unexamined the pluralism and misalignment of norms across overlapping reference networks within a broader social (or population) group under study.

The aim of this paper is to clarify these ambiguities by examining three aspects of Bicchieri's theory for connecting her theory to its methodological extensions for empirical norm modelling:[3]

1. the co-existence of four distinct kinds of social beliefs (first- and second-order, descriptive and normative),
2. the concept of belief consistency, understood in statistical rather than modal terms, and
3. the role of reference network alignment as a condition for norm identification.

By refining these aspects, I show that greater conceptual clarity directly enhances methodological practice in economics and the social sciences more broadly. The clarifications sharpen the empirical toolkit for modelling norms in experimental settings.

II. CO-EXISTENCE OF FOUR KINDS OF SOCIAL EXPECTATIONS/BELIEFS

The initial formal statement of the conditionality of social norms given by Bicchieri defines a social norm as a behavioural rule *R* for situations of type *S*, sustained when agents prefer to conform conditional on both empirical and normative expectations (Bicchieri 2006, 11). Her analysis thus places 'social expectations' at the core of norm identification.

Expectations, in her framework, are interchangeable with beliefs: they are about what will happen or what should happen (Bicchieri 2017, 11).

[2] I thank an anonymous reviewer for highlighting this point explicitly. Since reference networks are discussed throughout the paper, it is important to avoid a 'coextensive' reading that treats reference networks and social groups as the same.

[3] By 'empirical norm modelling', I primarily mean the operationalisation of Bicchieri's conceptual analysis for identifying the presence of a social norm. I recognise that the same framework may also be used to explain individual norm-compliant behaviour or to design norm interventions. The relation among these inferential aims is important, however, it remains as a separate issue outside the scope of this paper.

Bicchieri distinguishes *personal* beliefs—those held independently of others—from *social* beliefs, which are conditional on others' beliefs. Accordingly, norms are *personal* when grounded in individual conviction, and *social* when sustained by interdependent expectations. Her theory concerns only the latter, that is, the expectations we have about *others'* behaviours and beliefs.

Bicchieri identifies four types of social expectations (social beliefs):

1. First-order descriptive belief (1D): what one believes others do.
2. Second-order descriptive belief (2D): what one believes others believe I/others do.
3. First-order normative belief (1N): what one believes others should do.
4. Second-order normative belief (2N): what one believes others believe I/others should do.

TABLE 2.1 FULL SET OF SOCIAL BELIEFS REQUIRED IN BICCHIERI'S THEORY OF SOCIAL NORMS

SOCIAL BELIEF	FIRST-ORDER	SECOND-ORDER
DESCRIPTIVE	1D (condition 1) what one believes about what others do.	2D (condition 2) what one believes about what others believe I/others do.
NORMATIVE	1N (condition 3) what one believes about what others should do.	2N (condition 4)[4] what one believes about what others believe I/others should do.

Next, I will explain these *four* kinds of social beliefs one by one, relating to Bicchieri's original texts. I argue that *all* four types are conceptually necessary for norm identification. However, empirical applications citing Bicchieri and the ones done by Bicchieri and co-authors do not always preserve these distinctions in measurement practice or terminology.

[4] I am grateful to an anonymous referee for prompting me to clarify that Bicchieri's use of the label 'normative expectations' varies across works. In *The Grammar of Society* (2006), 'normative expectations' are introduced in terms of believing that others expect one to conform to a rule (a formulation closer to what I classify as second-order descriptive beliefs, 2D). In later work (for example, Bicchieri (2017)), 'normative expectations' are characterised more explicitly as beliefs about others' normative beliefs (closer to 2N). This terminological shift—and its interpretive implications—has also been noted in the secondary literature (for example, Guala (2017)). With this in mind, Table 2.1 should be read as an operational taxonomy intended to support empirical measurement and question design; it aligns primarily with Bicchieri's later usage and does not map onto her terminology in a one-to-one way.

II. I. EMPIRICAL EXPECTATIONS (1D AND 2D BELIEFS)

Conditions 1 and 2 in Bicchieri's framework concern empirical expectations.

The first-order descriptive belief (1D) is the belief about what others actually do. Bicchieri (2006, 12–13) links this to the *contingency condition*: a rule counts as a social norm only if agents are aware of it and expect others to conform. Without this awareness, there is no reason to adopt the rule as a guide to one's own behaviour. The second-order descriptive belief (2D) refers to what one believes others think I (or others) do. This condition ensures reciprocity: "agents expect others to follow it and […] they are expected by others to abide by the norm" (Bicchieri 2006, 112). In other words, 2D beliefs secure mutual anticipation and the adjustment of behaviour to what others believe about one's actions (Bicchieri 2006, 22).

In Bicchieri's original text, conditions 1 and 2 are differentiated by the concept of *conformity*. Bicchieri argues that

> to be effective, norms have to be activated by salient cues […]. A norm may exist, but it may not be followed simply because the relevant expectations are not there, or because one might be unaware of being in a situation to which the norm applies. Here the conditional preference for *conformity* refers to two senses of expectations that (a) agents expect others to follow it and (b) they believe that, in turn, they are expected by others to abide by the norm. (Bicchieri 2006, 112)

In this statement, (a) refers to the *contingency* condition—1D beliefs about what one believes others will do—whereas (b) refers to 2D beliefs about the 1D beliefs. Thus, according to Bicchieri, identifying social norms requires eliciting both 1D beliefs to test the *contingency* condition and 2D beliefs to test for *conformity*. This suggests a joint elicitation of 1D and 2D beliefs.

In Bicchieri's framework, 1D and 2D beliefs play distinct roles, but in empirical practice the label 'empirical expectations' is often used in ways that do not consistency tract this distinction. In fact, the expression 'empirical expectations' in her empirical application typically refers only to 1D beliefs (Bicchieri and Chavez 2010, 2013; Xiao and Bicchieri 2010; Bicchieri and Dimant 2022). However, the need to elicit 2D beliefs for norm identification is made explicit in Bicchieri's own work. For example, Bicchieri writes:

> To identify social norms, we usually only need to gauge personal normative beliefs and empirical/normative expectations, but on occasion we may also want to evaluate second-order empirical expectations, as when we ask people not just what they or others would do, but also what they think others believe people in their situation would (as opposed to should) do. (Bicchieri 2017, 70n11)

This passage may be read as allowing researchers to omit 2D measures in some settings. I emphasise that such omission should be treated as a context-specific modelling choice instead of a theoretical generalisation. As specified in Bicchieri (2006), a social norm requires that the action prescribed by the norm constitutes an equilibrium strategy. This is one reason second-order expectations (2D and 2N) play an important theoretical role in her account: they capture agents' beliefs about others' expectations, which are relevant to the mutual-adjustment structure that makes a pattern self-sustaining. Bicchieri characterises an equilibrium as a "situation of stable mutual adjustment: Everyone anticipates everyone else's behaviour, and these anticipations are met" (Bicchieri 2006, 22).

The methodological point here is not that we must always elicit 2D beliefs. Rather, it is that omitting 2D is an operational choice with inferential consequences. In many empirical designs (Bicchieri and Xiao 2009; d'Adda et al. 2020), 1D measures are sometimes used as a proxy for the expectation structure that (in the theory) is captured more directly by second-order descriptive beliefs. That proxy is most likely to fail when the relevant 'others' vary across reference networks, when compliance is driven by perceived social pressure (others' expectations) rather than by descriptive regularities, or when descriptive and second-order expectations come apart (for example, under pluralistic ignorance). These conditions are especially likely to arise in unfamiliar, asymmetric, or role-differentiated[5] situations, making the elicitation of 2D particularly important.

Consider two examples that illustrate when 1D measurement may be sufficient and when 2D should additionally be elicited. In a stable and familiar reference network, where the relevant behavioural equilibrium is close to public knowledge, 1D beliefs may provide a workable proxy for the expectation structure. For example, if students are asked what their

[5] See Witt (2023) for a discussion of the ontology of role-based normativity.

close friends would do in a familiar situation, their beliefs about what others do may be sufficiently informative for the limited purpose of testing whether behaviour tracks perceived regularities. By contrast, the 1D proxy is more likely to fail in a study of norms governing household relations between daughters-in-law and mothers-in-law in a society where such in-law relations are heterogeneous. It may not be sufficient to ask daughters-in-law what they think their mothers-in-law do, which elicits only 1D beliefs. It may also be necessary to elicit 2D beliefs, for example, by asking what they think their mothers-in-law expect them, as wives or daughters-in-law, to do. This is because, in practice, daughters-in-law may not be able to infer what they should do in a given situation simply by observing what their mothers-in-law do. Likewise, a mother-in-law may not infer the relevant norm simply by observing what daughters-in-law do, but by forming expectations about what daughters-in-law within that family or community expect her to do. In such cases, 2D beliefs are needed to capture the mutual expectation structure that sustains compliance.

Accordingly, the role of the equilibrium condition is to clarify why 2D is theoretically central, and why researchers should be explicit about when it is, and is not, safe to treat 1D-only measurement as sufficient.

II. II. NORMATIVE EXPECTATIONS (1N AND 2N BELIEFS)

Conditions 3 and 4 involve normative expectations.

Condition 3 refers to the first-order normative expectations (1N beliefs). In Bicchieri's analysis, she calls this type of belief *personal normative beliefs*. She further distinguishes these into two kinds. The first concerns beliefs about "what *I* should do"—the first-order *personal* normative beliefs in the strict sense. The second concerns beliefs about "what *others* should do"—the first-order *social* normative beliefs. Because Bicchieri's focus is on the social dimension, I set aside this internal distinction. Accordingly, in this paper my use of 1N belief refers only to the second type: *what one believes others should do*.

Condition 4 refers to the second-order normative expectations (2N beliefs). Bicchieri's analysis suggests a joint elicitation of 1N and 2N beliefs. On the one hand, Bicchieri (2017, 70) emphasises that 1N beliefs are needed to assess the mutual consistency between 1N and 2N beliefs. Such consistency arguably indicates the degree to which a behaviour is socially endorsed. For instance, an experimental study by Bicchieri (2022, 96) posed a 1N belief question to their subjects about child marriage: "Some

girls get married before they are 18 years old, is this good?". In the same study they posed a 2N belief question "Do you think people in your community believe that it is a father's duty to marry off daughters as soon as possible?". Bicchieri (2022) argues that if the 1N beliefs and the 2N beliefs do not align, the accuracy of the normative expectations (2N beliefs) cannot be properly assessed. This underscores the need for their joint elicitation.

On the other hand, Bicchieri stresses that 2N beliefs are required to test 1N beliefs. According to her analysis, 2N beliefs provide crucial evidence that behaviour is governed by a norm rather than by individual preference. As Bicchieri (2006, 15) explains, people may have varied reasons for conditionally preferring to follow a rule. If an agent unconditionally behaves in accordance with a shared belief, it is likely that the agent is following a *personal* normative belief (for example, a prudential norm),[6] where practical reasoning weighs more heavily than social constraint. By contrast, if an agent appeals to a shared belief (attitude) precisely because of social expectations, then practical reasoning is less decisive, and compliance is better explained by the presence of a social norm.

Although Bicchieri's theoretical distinction between 1N and 2N beliefs is clear, empirical operationalisation (including some by Bicchieri and coauthors) sometimes primarily focus on 2N measures, leaving 1N unmeasured or implicit. This choice should be considered methodological: an inferential choice that affects what the study can establish about norm endorsement and compliance. For example, consider a practice that many individuals privately judge wrong (low 1N) while believing that most others approve of it (high 2N). A 2N-only operationalisation would register strong "normative expectations", whereas joint elicitation of 1N and 2N would reveal divergence between personal endorsements and perceived social endorsements—an empirically important pattern for diagnosing phenomena modelled as pluralistic ignorance or preference falsification.[7] This is not to deny that 2N-only measures can be appropriate for some

[6] A prudential norm is what agents comply with due to prudential reasoning (prudential belief); it is regarded as unconditional or less conditional. In Bicchieri's analysis, she disregards prudential and moral norms in her theory of social norms, although she claims that they are *social* in the weak sense (2017, 71). The 'weak sense' is what I mean here 'less conditional'.

[7] I am grateful to one anonymous reviewer for highlighting this issue. In a separate work (Ross and Wang 2025), I examine how Bicchieri's account can be reconciled with Kuran's account of preference falsification, and I clarify the distinction between pluralistic ignorance and preference falsification as it bears on Bicchieri's framework. This issue is also noted later in Section 3.1, however, given the focus and space constraints of the present paper, I set this issue aside here.

research aims. The point is that omitting 1N narrows the inferences a design can support—especially in contexts where measuring 1N is important for distinguishing internalisation from perceived social pressure.

Though Bicchieri (2017, 70) suggests in most empirical cases to treat 1N beliefs as secondary, her conceptual framework makes room for both components. I therefore argue that distinguishing 1N and 2N beliefs is conceptually important, and that eliciting both or either should be left as an inferential choice.

II. III. CLARIFYING THE CONCEPTUAL NECESSITY OF ALL FOUR BELIEF TYPES

The ambiguity I have noted here may create the impression that some empirical applications—including some studies by Bicchieri and co-authors—treat 1D and 2N beliefs as especially diagnostic for norm identification, while leaving 2D and 1N beliefs less/un- measured.[8] However, I maintain that this is an *empirical* and inferential issue, not a conceptual one.[9] The relative importance of different belief types should be determined by the specific social contexts in which norms operate, rather than by excluding some types *a priori*. Accordingly, when studies elicit only 1D and 2N beliefs, that choice should be made explicit and justified in relation to the intended inference, and—where appropriate—examined against alternative elicitation sets.

I do not deny the possibility that some kind of beliefs may carry less weight depending on the empirical circumstances. This conjecture is captured in Kuran's (1997) model of social influences on utility, which attaches weighting parameters to different types of beliefs. Bicchieri's model has no comparable resource. Thus, there is no basis in applying Bicchieri's conceptual framework for social norms to exclude any one of the four types of social beliefs *a priori*. For conceptual clarity, the full taxonomy should be made explicit, while empirical studies may measure subsets with justifications.

To conclude, this section clarified the ambiguity regarding one set of conditions under which social norms can be modelled empirically within Bicchieri's framework. I have argued that norm identification requires the co-existence of four distinct kinds of social expectations—captured by

[8] For example, in Bicchieri and Chavez (2010) and Bicchieri and Dimant (2022) only 1D and 2N beliefs are elicited.

[9] I am grateful for one anonymous reviewer's suggested that "it depends on what researchers are looking for", and I agree.

1D, 2D, 1N, and 2N beliefs. While some empirical applications, including studies by Bicchieri and co-authors, may give greater operational weight to certain belief types, I contend that all four remain conceptually relevant and should be explicitly acknowledged. The implication for empirical research is not that all four belief types must always be elicited, but that the selected elicitation set should be justified in advance in relation to the target norm, the relevant social context, and the intended inference.

III. BELIEF CONSISTENCY

Corresponding to the four co-existing kinds of social belief, a second aspect in Bicchieri's theory of norm identification is *belief consistency*. However, there is some ambiguity about what belief consistency means. In this section I clarify the concept and then illustrate potential implications for statistical tests of observed choice behaviour to identify norms.

III. I. DEFINING BELIEF CONSISTENCY

What does it mean to say that beliefs are consistent? How similar must they be to count as consistent? From a statistical point of view, how much overlap between two belief-response distributions (whether elicited at individual level[10] or observed as the distribution of point responses within a reference network) satisfies a threshold for consistency? What is required, with respect to central tendency (means or modes) and dispersion (for example, standard deviations), to conclude they are consistent?

The notion of similarity has been analysed from a feature-theoretical approach by Tversky (1977), which challenges the previous domination of geometric models. Geometric models measure the similarity relation as the "metric distance between points in coordination space" (Tversky and Gati 1978, 75). However, a feature-theoretical approach defines similarity between objects as the value of a feature-matching function. Based on quantitative assumptions about similarity ordering, the similarity of two objects is characterised as "a linear combination (or contrast) of the measures of their common and distinctive features" (Tversky and Gati 1978, 76). Similarity increases with the measure of commonality and decreases with the measure of distinction. Tversky and Gati (1978) employed the contrast model to analyse three empirical problems. Their

[10] For the rationale behind eliciting beliefs as distributions in norm modelling within Bicchieri's framework, see Wang (2023).

results suggest that the parameters of the contrast model are sensitive to manipulation.

In Bicchieri's theory of social norms, there is a distinction between *across-individual* consistency and *within-individual* consistency.

Across-individual belief consistency concerns the degree of consistency of a belief across individuals in a community (reference network). It is what Bicchieri calls mutual consistency or *consensus*. Statistically, this kind of consistency is modelled by aggregated data at the level of society (group-level). For example, Bicchieri (2017, 70) argues that "when a social norm exists and applies to a specific situation, the second-order *normative expectations* (between different individuals) will be mutually consistent. If these expectations are mutually consistent, then there is widespread *consensus* that a specific behaviour should be performed (or avoided)". In other words, "consensus means an agreement in *individuals' second-order normative expectations,* that is the presence of a shared norm" (Bicchieri and Chavez 2010, 163).[11] This refers to the consistency of the 2N beliefs.

By contrast, within-individual belief consistency concerns the degree of consistency among different but relevant beliefs held by an individual. In her discussion of pluralistic ignorance, Bicchieri (2017, 74) argues that "when second-order normative beliefs (2N beliefs) are mutually consistent but systematically inaccurate, we know that people uphold a norm they dislike". This diagnose involves a comparison between across-individual consistency with within-individual consistency. 'Mutually consistent' refers to the across-individual convergence of 2N beliefs within the group, whereas 'systematically inaccurate' refers to a mismatch between individuals' 2N beliefs and the actual distribution of 1N beliefs in the relevant reference network. Such an application takes the standpoint

[11] Bicchieri and Chavez's (2010) study illustrates why the statistical form of 'agreement' matters. In their modified Ultimatum Game, they report "a remarkable degree of agreement" between Responders' first-order normative expectations about fairness and Proposers' beliefs about those expectations. On the interpretation developed here, this finding establishes agreement between selected point-estimate summaries across roles and levels of belief, but not yet distribution-sensitive across-level consistency. To establish the stronger conclusion, one would need to know whether the underlying belief distributions are similarly shaped, sufficiently concentrated, and not masking dispersion, polarisation, or subgroup-specific disagreement. The same mean-level agreement could reflect tight convergence, but it could also arise from heterogeneous beliefs that average out at the aggregate level. The implication is not that Bicchieri and Chavez's (2010) behavioural findings are mistaken, but that the norm-identification claim should be qualified: their study provides evidence of aggregate point-estimate agreement in selected normative components, while the stronger claim of robust belief consistency depends on the distributional profile of the underlying beliefs.

of individuals in explaining individual norm-compliant behaviours, which together contributes to the collective phenomenon of pluralistic ignorance.[12]

This distinction matters for how belief consistency is treated in empirical studies implementing Bicchieri's account. Individual belief responses are often interpreted through whole-sample or community-level summaries. Such an approach may be appropriate for the narrower aim of testing whether respondents' perceived social expectations are associated with, or predict, their own behaviour, however it may not be sufficient for the identification of the presence of a norm, since the relevant reference network (Section 4) may not coincide with the whole sample. Bicchieri et al.'s (2022) study of exclusive breastfeeding in Mali illustrates the point. They examine whether mothers' perceptions of community behaviour and community approval predict their own breastfeeding behaviour. They define 'community' as "people who live in your area, such as family, friends and neighbours", allowing respondents to refer broadly to their own reference network. This is a reasonable strategy for testing whether perceived social expectations are behaviourally relevant. However, for norm identification, further specification may be needed. A young mother's relevant network may include husbands, mothers-in-law, older women, health workers, neighbours, or other mothers, and these groups may hold different expectations. A broad 'community' category may therefore obscure whose expectations matter most, and whether the relevant expectations are sufficiently shared within the network that regulates the practice.

A natural extension, then, is to connect individual-level consistency profiles with reference-network-level consistency measures. In typical experimental settings operationalising Bicchieri's model for norm identification, individual beliefs of each subject will be elicited. Researchers can first examine whether each respondent's social expectations are internally consistent, and whether these expectations jointly support norm compliance. These individual-level consistency profiles can then be compared across respondents to ask whether they form meaningful subgroup or reference-network clusters. One could technically ask if there exist arbitrary sets of individuals of a given size for which the hypothesis of a social norm is accepted, but that is not a particularly interesting question.

[12] Applying such methodology confuses pluralistic ignorance with prevalent social phenomena modelled as preference falsification by Kuran (1997), which I have argued elsewhere (Ross and Wang 2025). Nevertheless, given the goal of this paper, I will leave this discussion aside.

A more interesting question would be the evaluation of stratified subgroups in terms of observed demographic characteristics (such as men or women, or racial groups) or the responses to a values survey (for example, attitudes towards the role of government), to determine whether subgroups display beliefs consistent with the presence of a social norm.

III. II. STATISTICAL INSIGHTS

The preceding discussion shows that belief consistency is not a single relation. It may concern the relation among an individual's own beliefs, the convergence of beliefs across members of a reference network, or the relation between individual-level belief profiles and group-level distributions. Once belief consistency is understood as one conceptual condition for norm identification in Bicchieri's account, the next question asks how it can be operationalised. Answering this requires specifying a *statistical* criterion—a decision rule based on observed proportions together with information about dispersion or polarisation in the response distribution. Importantly, this issue arises regardless of whether beliefs are elicited as full distributions or as point estimates:[13] point responses still generate empirical distributions at the group (or reference-network) level, and different distributional profiles can support different inferences even when a simple majority criterion is met. where feasible, eliciting belief distributions can make these patterns more directly observable.

Three issues require clarification.

First, which belief types should be assessed for consistency?

Should *all* four types of beliefs be modelled, or is it sufficient to measure only certain types of beliefs? In current empirical work, it is common practice that only some belief types are elicited. For example, the empirical research by Bicchieri and co-authors followed such a practice: Bicchieri and Chavez (2010) elicited 1D and 2N beliefs; Xiao and Bicchieri (2010) elicited 1D and 1N beliefs; and Bicchieri and Dimant (2022) elicited 1D and 2N beliefs. Empirical work done by other scholars followed similar methods (see Pérez León 2022; Khalmetski 2016). This pattern suggests that some belief types are elicited more frequently than others in current

[13] In related work, I argue that eliciting belief distributions is often methodologically preferable for implementing Bicchieri-style norm diagnostics (Wang 2023). For the foundational scoring-rule method for eliciting subjective belief distributions, see Harrison et al. (2017). The present discussion is compatible with that recommendation but does not presuppose it: the same inferential issues arise when beliefs are elicited as point estimates, since point responses still generate an empirical distribution at the reference-network level.

empirical work. In particular, 1D and 2N beliefs are often elicited, while 2D and 1N are less often measured. Consistent with Section 2, however, I sustain that this should be treated as an *empirical* rather than a conceptual matter: researchers should decide, given the social context of the target norm and the intended inference, which belief types are necessary for assessing consistency. Accordingly, measurement should first determine which social beliefs are relevant in context and then decide whether to elicit a *subset* or *all* four types for consistency tests.[14]

Second, what counts as consistency for a 'majority' threshold?

This issue concerns the *scale* and *decision rule* used to assess consistency at the group (or reference-network) level. Bicchieri (2017, 42, 62, 69, etc.) characterises belief consistency in terms of alignment between an individual's expectations with what a *majority* in their reference network[15] approve or disapprove of. Yet what qualifies as a 'majority' is not fixed: it can vary across people, across norms and settings, and Bicchieri herself notes that the required proportion differs case by case (2017, 108). Accordingly, empirical applications should treat the majority threshold as an explicit modelling choice—for example, a threshold q on the proportion p of responses in the focal category (or within a tolerance band)—rather than an implicit default (such as 50%). A distribution-sensitive approach then should report not only whether $p \geq q$, but also the distribution profile of responses (for example, concentration versus dispersion or polarisation), since the same value of p can reflect either tight clustering near the agreement category (near the consensus) or sharp minority disagreement that indicates contested expectations.[16]

This issue is tied to reference networks (next section). Reference networks define sub-group norms, because reference networks determine normative expectations and what actions and attitudes people care about. What constitutes a reference network depends on the interactions people have within a community.

Third, how to treat overall versus partial consistency?

---

[14] I thank one anonymous reviewer for prompting me to make it explicit that which beliefs matter for consistency is target-relative.

[15] Here 'majority' should be understood relative to the relevant reference network *Ni* rather than the entire sampled population, unless the design explicitly justifies treating the two as aligned.

[16] The point generalises beyond categorical coding. Even when a belief is elicited on a graded scale (such as Likert scale), reliance on a single summary can obscure structure: groups can look similar on one aggregate statistic while differing substantially in dispersion or multimodality. In parametric terms, two samples may share s similar mean (or mode) yet differ in variance; in nonparametric terms, they may differ in tail mass or bimodality.

Before distinguishing overall and partial consistency, it is helpful to clarify what 'consistency' is defined over. Consistency may be assessed within individuals (coherence across a single person's 1D, 2D, 1N, and 2N beliefs) at the group or reference-network level (the extent to which individuals' beliefs align with one another), and component-wise (whether alignment is stronger for some belief components or relations than for others). These senses can come apart empirically, so the analysis should specify the relevant level of assessment in each case.

With this in mind, let *overall* belief consistency refer to the degree of alignment across all belief types elicited (i.e., 1D, 2D, 1N, and 2N beliefs) at a specified level of analysis (within a reference network). *Partial* consistency refers to empirical phenomena in which alignment holds for some components or relations but not others—for example, descriptive components may be mutually aligned (for example, 1D with 2D) while normative components are not (for example, 1N with 2N), or alignment may hold for a subset of relations but break down elsewhere. The methodological point is that empirical evidence often exhibits such mixed patterns: treating consistency as an all-or-nothing property risks obscuring where agreement is concentrated and where it is contested.

In experimental settings, tests of overall consistency across all belief types can be extremely complex, both conceptually and statistically. A useful strategy is therefore to analyse partial-consistency components (for example, selected relations among belief types or outcome categories) and to use these results to construct or motivate an overall assessment via an explicit aggregation rule. In statistical practice, an *overall* consistency across *all four* types of belief can result in various levels of partial consistency, for example, it can be supported by multiple component tests that can yield different degrees of alignment. For example, in the experimental study run by Wang (2022), the primary goal was to test the overall belief consistency across all four types of beliefs. I started with testing component relations (partial consistency)—between 1D and 1N beliefs, and between 2D and 2N beliefs, and then used these component results to characterise the degree of overall consistency at the group level. More generally, one can represent overall belief consistency as a weighted aggregation of partial consistencies across the relevant components (with one possible implementation being a weighted product across outcome categories). Making this relationship explicit allows partial-consistency patterns to be reported rather than being absorbed into a binary verdict.

The distinction between overall consistency and partial consistency opens the possibility of variations of degrees of consistency being identified by demographic variables that partition groups into different reference networks. Partial belief consistency across different reference networks is what causes norm uncertainty. Statistical modelling of belief inconsistency in this sense should be able to explain what degree of consistency is required for norm compliance, and what degrees of inconsistency can explain social phenomena such as the existence of norm disruptors, norm manipulation, and pluralistic ignorance. For example, Dana, Weber, and Kuang (2007) conducted an experimental study in which they gave subjects moral 'wiggle room' in interpreting norm compliance and found that subjects given such scope for rationalisation behaved more self-interestedly. Tracking inconsistency between empirical and normative beliefs in similar experimental settings may yield inferences which explain why fairness is a social norm rather than a set of social preferences.

In sum, statistical modelling of belief consistency requires careful attention to which belief types are compared, how majority thresholds are defined within reference networks, and how overall versus partial consistency is assessed, since these factors jointly determine the strength and scope of norm identification.

IV. REFERENCE NETWORK

IV. I. DEFINING A REFERENCE NETWORK

Reference network is the third aspect of Bicchieri's theory for norm identification. Bicchieri (2017, 53) characterises a reference network N_i as "the network of people whose behavior and expectations matter" to an individual agent i. A reference network is a cluster within a larger social influence network (Goyal 2009).[17] Importantly, a social group (the population under study) need not coincide with any individual's reference network: reference networks are typically proper subsets, may overlap, and can differ across individual and contexts. Here, *clustering* simply means that the density of mind shaping relationships[18] is high within the cluster relative to other parts of the larger network (Zawidzki 2013).

Section 3.1 noted that experimental implementations of Bicchieri's account (for example, Bicchieri and Chavez 2010; d'Adda et al. 2020)

[17] See Goyal (2009, chapter 2) for the relevant concept specification.
[18] See Zawidzki (2013).

frequently operationalise belief consistency at the aggregate sample level, implicitly assuming a homogeneity of the sample group. The assumption of homogeneity can obscure the role of reference networks in applying Bicchieri's account for norm modelling. In many contexts, the behaviourally relevant reference network is often a sub-group (or cross-cutting network), so using the full sample as the unit of comparison may mask substantially important heterogeneity.

Reference networks may involve heterogeneity across several dimensions. First, scale: a reference network can be local or general, and be big or small. Second, conditions of activation: different contexts can determine when a reference network becomes salient for an agent's decision making. Third, dynamics: reference networks can change over time. The very concept of a reference network reflects that a norm is always community-specific, where the 'community' may be defined broadly or narrowly.

According to Bicchieri's analysis, reference networks matter for shaping agents' conditional preferences, particularly the 2N beliefs.[19] Bicchieri (2017, 53) warns that ignoring reference networks in a normative intervention is likely to lead to failure. One example is a failed intervention to promote breastfeeding in West and Central Africa. In the study, researchers (Ayoya et al. 2012) attempted to change behaviours of new mothers, to encourage them to breastfeed their babies. The intervention consisted of information provision. As a result, behaviours of mothers did not significantly change. The failure arguably stemmed from the fact that mothers-in-law and older women in general, who play a key role in young mothers' reference networks, were not included in the intervention. Unchanged social expectations, particularly the normative expectations within those reference networks, outweighed the informational treatment.

The heterogeneity of reference networks also reflects the fact that individuals live in interacting social webs. Agents' decisions in different situations may be shaped by different reference networks, and as circumstances change, the relative influence of members of those networks may also shift. Different reference networks may carry different normative expectations, yet only the expectations of those who *matter* motivate norm compliance or violation. Bicchieri (2017, 14) illustrates this with the case of a Pakistani immigrant to Italy who killed his 'dishonoured' daughter even though he had lived in Milan for 20 years. The social norms

[19] For discussion of the importance of normative expectations, see Bicchieri and Sontuoso (2020) and Wang, Juzaszek, and Ryan (2025).

prevailing in Milan and the normative expectations of his friends and colleagues in Milan were not as effective in influencing him as the norm of family honour prevalent in his Pakistani village. The family and people in the village played the role of the reference network governing his response to his daughter's perceived transgression.

Finally, reference networks are dynamic. Muldoon (2022) argues that when reference networks split, some sub-groups stop following previous norms. This suggests that one potential approach to norm intervention is to alter reference networks, which in turn can change agents' social expectations, especially normative expectations. Bicchieri (2017, 158) argues that if deliberation is used to change a norm, it should not only target the 'at-risk' group (such as young women, in the breastfeeding case), but must include all members of the relevant reference network—such as men, relatives, extended families, religious and village leaders—whose expectations may change or realign over time.

This highlights why reference networks must be made explicit not only in conceptual clarification but also in empirical modelling of norm dynamics.

IV. II. METHODOLOGICAL AND STATISTICAL INSIGHTS

Norm modelling must be sensitive to reference networks in the sense that any investigation of a social norm must specify the population (social group) under study and, within it, the relevant reference networks. Empirical research should clearly identify which reference network is relevant when eliciting subjects' beliefs. Such networks can sometimes be inferred through demographic and/or normative surveys. Moreover, providing subjects with information about reference networks may prime them to frame their responses in relation to that network. In statistical modelling, such information can be incorporated as covariates. Empirical studies may also measure the degree of consistency across different reference networks by comparing demographic subgroups in samples and across treatments.

With respect to the scale of a reference network, what matters to empirical studies is agents' *confidence levels* in their beliefs (Harrison and Swarthout 2020). Suppose a study randomly sampled subjects across a very large community, across a nation, for example. We might expect that any social norm we found would be reflected in low confidence across subjects' beliefs. This is because subjects might anticipate significant heterogeneity across a whole nation. This would be expected to lead to

diffuse prior beliefs. By contrast, if a study used a highly cohesive community sample with a tight reference network, and if subjects were aware of the reference network, this would be expected to lead to higher confidence in beliefs.

The dynamics of the reference networks are equally important: in studies conducted over an extended period, experimenters should allow for the possibility that subjects' reference networks may shift over time. The design of the study therefore should capture corresponding changes in subjects' beliefs over the study's time course.

Finally, empirical identification and control of reference networks can be achieved through demographic and normative surveys, which provide exogenous information about the networks most relevant to subjects' beliefs. A practical approach is to collect demographic information (such as gender, age, education) and normative attitudes. For example, in the experiment conducted by Wang (2022) in South Africa, reference network alignment was controlled by administering a demographic questionnaire, and a normative values survey which was developed on the basis of a selection of from Afrobarometer items.[20] The aim was to obtain exogenous information to better control for a normative reference network across subjects.

In summary, when adopting Bicchieri's model for empirical norm modelling, the first step must be experimental control for reference network alignment. If there exists a social norm of relevant expectations/beliefs that align, the simplest case for application of Bicchieri's theory arises where we expect the norms of interest (i.e. behaviour in Trust Game), in *one* reference network shared by everyone. In fact, in her discussion of measuring and changing norms, Bicchieri emphasises the same procedure that "mapping the reference network is an essential part of understanding social norms and how to change them, because the norm has to change within the reference network" (2017, 53). In cases where there are multiple reference networks at play, we must expect consequent complexity of normative behaviour. In particular, we should expect to see preference falsification (Kuran 1997) or pluralistic ignorance, due to existence of belief inconsistency across mis-aligned reference networks.

[20] Afrobarometer questionnaires are applied widely across Africa to explore the breadth of African public opinions. More information can be found at https://www.afrobarometer.org/surveys-and-methods/

IV. III. NORM PLURALISM

A further implication of the reference network framework is that it highlights the phenomenon of *norm pluralism*. Because individuals typically belong to multiple, overlapping reference network clusters (within the same wider social group) may sustain different, even conflicting, social expectations. This pluralism can be understood in equilibrium terms: a given community may coordinate on one equilibrium (such as a fairness norm in market exchange), while another community coordinates on a different equilibrium (such as a family-honour norm). The existence of such pluralism raises philosophical questions about the ontology of norms. Are norms relative to reference networks, so that multiple norms can co-exist without contradiction, or should norms be conceived as universal social facts that hold across groups? Bicchieri's framework, once clarified in terms of reference networks, suggests its application to the former: norms are best understood as cluster-relative equilibria sustained by particular patterns of mutual expectations. This view resonates with Kuran's (1997) analysis of preference falsification, which shows how individuals may publicly align with one network's norms while privately endorsing another. It also aligns with Sugden's (1998, 2004) account of conventions as local solutions to coordination problems, and with Binmore's (1994, 1998) claim that different equilibria can emerge as focal points in different communities. Taken together, these perspectives indicate that the operational task of mapping reference networks is directly tied to the philosophical problem of norm pluralism: how to reconcile normative diversity within and across societies.

V. CONCLUSION

To conclude, this paper has developed an operational refinement of Bicchieri's framework for empirical norm modelling by clarifying three elements: (1) the taxonomy of norm-relevant beliefs, (2) the assessment of belief consistency, and (3) the specification and alignment of reference networks. In clarifying each aspect, I identified sources of ambiguity in how the framework is operationalised, drew out methodological and statistical implications.

The refinements offered here focus on the operational objectives of Bicchieri's framework in empirical applications. On belief taxonomy, I argued that it should be acknowledged at the conceptual level, while leaving it as an empirical and inferential question whether a subset or the complete set should be modelled in specific contexts. Second, regarding belief

consistency, I clarified that Bicchieri's model is explicit about across-individual consistency at the group-level, while also noting its potential to incorporate within-individual consistency. I further highlighted three elements that empirical applications must consider when modelling belief consistency: the choice of belief types, the specification of majority thresholds, and the distinction between overall and partial consistency. Third, regarding reference networks, I emphasised their heterogeneity in terms of scale, activation conditions, and dynamics, and stressed the importance of explicit specification—and where feasible, measurement or control of reference-network alignment when applying the framework empirically.

REFERENCES

Andrews, Kristin. 2020. "Naïve Normativity: The Social Foundation of Moral Cognition." *Journal of the American Philosophical Association* 6 (1): 36–56.

Ayoya, Mohamed A., Malek A. Bendech, Noel M. Zagré, and Felicité Tchibindat. 2012. "Maternal Anaemia in West and Central Africa: Time for Urgent Action." *Public Health Nutrition* 15 (5): 916–27.

Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, Cristina. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Bicchieri, Cristina. 2022. "Norm Nudging: How to Measure What We Want to Implement." In *Behavioral Science in the Wild*, edited by Nina Mazar and Dilip Soman, 82–107. University of Toronto Press.

Bicchieri, Cristina, and Alex K. Chavez. 2010. "Behaving as Expected: Public Information and Fairness Norms." *Journal of Behavioral Decision Making* 23 (2): 161–78.

Bicchieri, Cristina, and Alex K. Chavez. 2013. "Norm Manipulation, Norm Evasion: Experimental Evidence." *Economics & Philosophy* 29 (2): 175–98.

Bicchieri, Cristina, and Eugen Dimant. 2022. "Nudging with Care: The Risks and Benefits of Social Information." *Public Choice* 191 (3–4): 443–64.

Bicchieri, Cristina, and Erte Xiao. 2009. "Do the right thing: but only if others do so." *Journal of Behavioral Decision Making* 22 (2): 191-208.

Bicchieri, Cristina, Erte Xiao, and Ryan Muldoon. 2011. "Trustworthiness Is a Social Norm, but Trusting Is Not." *Politics, Philosophy & Economics* 10 (2): 170–87.

Bicchieri, Cristina, and Jingnan Zhang. 2012. "An Embarrassment of Riches: Modeling Social Preferences in Ultimatum Games." In *Games, Rationality and Behaviour: Essays on Behavioural Game Theory and Experiments*, edited by Alessandro Innocenti and Patrizia Sbriglia, 97–122. Palgrave Macmillan.

Bicchieri, Cristina, and Alessandro Sontuoso. 2020."Game-theoretic accounts of social norms: The role of normative expectations." In *Handbook of experimental game theory*, edited by C. Mónica Capra, Rachel T. A. Croson, Mary L. Rigdon, and Tanya S. Rosenblat, 241-255. Edward Elgar Publishing.

Bicchieri, Cristina, Urmi Das, Sarah Gant, and Rebecca Sander. 2022. "Examining Norms and Social Expectations Surrounding Exclusive Breastfeeding: Evidence from Mali." *World Development* 153: 105824. https://doi.org/10.1016/j.worlddev.2022.105824

Binmore, Kenneth G. 1994. *Game Theory and the Social Contract, Vol. 1: Playing Fair*. MIT Press.

Binmore, Kenneth G. 1998. *Game Theory and the Social Contract, Vol. 2: Just Playing*. MIT Press.

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood. 2013. *Explaining Norms*. Oxford University Press.

d'Adda, Giovanna, Martin Dufwenberg, Francesco Passarelli, and Guido Tabellini. 2020. "Social norms with private values: Theory and experiments." *Games and Economic Behavior* 124: 288-304.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33: 67–80.

Dennett, Daniel C. 1987. *The Intentional Stance*. MIT Press.

Goyal, Sanjeev. 2012. *Connections: An Introduction to the Economics of Networks*. Princeton University Press.

Guala, Francesco. 2007. "Grammar of Society: The Nature and Dynamics of Social Norms, by Cristina Bicchieri." *The British Journal for the Philosophy of Science* 58 (3): 613–18.

Guala, Francesco. 2017. "Review of Cristina Bicchieri's *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*." *Erasmus Journal for Philosophy and Economics* 10 (1): 101–11.

Harrison, Glenn W., Jimmy Martínez-Correa, J. Todd Swarthout, and Eric R. Ulm. 2017. "Scoring rules for subjective probability distributions." *Journal of Economic Behavior & Organization* 134: 430-448.

Harrison, Glenn W., and J. Todd Swarthout. 2020. *Belief distributions, Bayes rule and Appropriate Confidence*. Vol. 11. Working Paper 2020.

House, Bailey R., and Michael Tomasello. 2018. "Modeling Social Norms Increases Costly Sharing in Middle Childhood." *Journal of Experimental Child Psychology* 171: 84–98.

Huber, Christoph, Christos Litsios, Andreas Nieper, and Tobias Promann. 2023. "On Social Norms and Observability in (Dis)honest Behavior." *Journal of Economic Behavior & Organization* 212: 1086–99.

Khalmetski, Kiryl. 2016. "Testing Guilt Aversion with an Exogenous Shift in Beliefs." *Games and Economic Behavior* 97: 110–19.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Muldoon, Ryan. 2022. "Social Norms." In *The Routledge Handbook of Philosophy, Politics, and Economics*, edited by Chris M. Melenovsky, 137–48. Routledge.

Pérez León, Rodrigo. 2022. "Why Mexicans (Dis)Obey the Law: Social Norms, Legal Punishment and Pluralistic Ignorance." *Public Governance, Administration and Finances Law Review* 7 (2): 91–108.

Peterburs, Jutta, Robert Voegler, Robin Liepelt, André Schulze, Sabine Wilhelm, Sebastian Ocklenburg, and Thomas Straube. 2017. "Processing of Fair and Unfair Offers in the Ultimatum Game under Social Observation." *Scientific Reports* 7 (1): 44062.

Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, Don. 2014. *Philosophy of Economics*. Palgrave Philosophy Today. Palgrave Macmillan.

Ross, Don. 2022. "Neo-Samuelsonian Methodology, Normative Economics, and the Quantitative Intentional Stance." In *Methodology and History of Economics: Reflections with and without Rules*, edited by Bruce Caldwell, John Davis, Uskali Mäki, and Esther-Mirjam Sent, 90–116. Routledge.

Ross, Don, and Cuizhu Wang. 2025. "Modelling Conditionally Respected Social Norms: A Critique from the Intentional Stance." *Journal of Economic Methodology*. https://doi.org/10.1080/1350178X.2025.2535369

Ross, Don, Wayne C. Stirling, and Luca Tummolini. 2021. *Modeling Norm-Governed Communities with Conditional Games*. Working Paper WP 2021-05. Georgia State University, Center for the Economic Analysis of Risk.

Sugden, Robert. 1998. "Normative Expectations: The Simultaneous Evolution of Institutions and Norms." In *Economics, Values, and Organization*, edited by Avner Ben-Ner and Louis Putterman, 73–100. Cambridge University Press.

Sugden, Robert. 2004. *The Economics of Rights, Co-operation, and Welfare*. Palgrave Macmillan.

Tversky, Amos. 1977. "Features of Similarity." *Psychological Review* 84 (4): 327–52.

Tversky, Amos, and Itamar Gati. 1978. "Studies of Similarity." In *Cognition and Categorization*, edited by Eleanor Rosch and Barbara B. Lloyd, 79–98. Lawrence Erlbaum.

Vriens, Eva, Luca Tummolini, and Guido Andrighetto. 2023. "Vaccine-Hesitant People Misperceive the Social Norm of Vaccination." *PNAS Nexus* 2: 1–11.

Wang, Cuizhu. 2022. "Normative Expectations and Subjective Beliefs: An Incentivised Experimental Study." *PhD thesis*, University College Cork.

Wang, Cuizhu. 2023. "Belief Distributions and the Measure of Social Norms." In *Experiments in Moral and Political Philosophy*, edited by Hugo Viciana, Antonio Gaitán, and Fernando Aguiar, 125–43. Routledge.

Wang, Cuizhu, Maciej Juzaszek, and Shane Ryan. 2025. "Epistemic Nudging for Accurate Normative Expectations." *Synthese* 206 (1): Article 54. https://doi.org/10.1007/s11229-025-05127-y.

Witt, Charlotte. 2023. *Social Goodness: The Ontology of Social Norms*. Oxford University Press. ISBN 9780197574799.

Xiao, Erte, and Cristina Bicchieri. 2010. "When Equality Trumps Reciprocity." *Journal of Economic Psychology* 31 (3): 456–70.

Zawidzki, Tadeusz. 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.

Cuizhu Wang is a Research Fellow at Trinity Business School, Trinity College Dublin. Her research lies at the intersection of social philosophy, behavioural economics, and the philosophy and methodology of the social sciences. She works on social norms, moral psychology, belief modelling, and decision modelling, with a particular interest in how norm-related beliefs can be measured and modelled empirically. Her work combines theoretical analysis with experimental and methodological approaches. Cuizhu's research is funded by the European Union (ERC-2022-STG, CMP,

101077471). View and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.
Contact e-mail: <dawncz_wang@hotmail.com>