nature human behaviour

Perspective

https://doi.org/10.1038/s41562-026-02557-7

An agenda for real-world fairness research to inform policy and organizations

Received: 8 January 2026

Accepted: 17 July 2026

Published online: 17 August 2026

Check for updates

Ernst Fehr [1,3] & Keyu Wu [2,3]

Fairness concerns shape how people judge wages, institutions, taxes, transfers and social conflict. Decades of research show that people care not only about their own outcomes but also about how outcomes are distributed across others. The key question is now how these concerns operate outside simple laboratory settings, in real social and institutional environments. We argue that fairness concerns are shaped by comparison groups, beliefs about what outcomes are deserved, social norms, early-life experience and the institutions that make some comparisons more visible than others. This means that policies and organizations do not merely respond to fairness preferences; they can also activate, redirect and reshape them. Progress requires methods that connect experiments with field data, institutional variation and richer measures of how people evaluate whole income distributions. Such work can improve the design of workplaces, welfare policies and public institutions that are both effective and seen as legitimate.

Why do workers reduce effort after wage cuts, citizens punish free-riders at their own expense, affluent voters support redistribution from rich to poor groups although this reduces their net income, and people sometimes even risk their lives to protest political and institutional injustice? A common thread running through these diverse behaviours is concern for fairness.

A large body of behavioural evidence shows that individuals evaluate not only their own outcomes but also the outcomes and behaviour of others, including whether these are perceived as fair, efficient, kind or deserved[1,2]. These social preferences (see Box 1 for a glossary of key concepts) are also supported by evidence from neuroscience showing that neural activity in specific brain circuits, and experimental modulation of these circuits, can strengthen or weaken fairness-driven behaviour by influencing, for example, rejection rates of responders in ultimatum games or donations in dictator games[3]. Thus, the first-generation question of whether people have behaviourally meaningful preferences for fairness has largely been settled.

This foundational progress reshapes which questions should now stand at the centre of inquiry. In this Perspective, we argue that the pressing frontier is now application-oriented: fairness concerns do not operate in a vacuum; they operate through institutions that determine which comparisons become salient, which outcomes appear deserved, how norms and preferences are formed, and how fairness concerns are scaled to whole income distributions.

This shift is not simply a move from basic laboratory research to applied research. That is, institutional settings not only allow researchers to test laboratory insights for external validity, but they also reveal mechanisms that stylized experiments often hold fixed, including reference-group selection, procedural legitimacy, norm internalization and the activation of different fairness ideals. Studying these processes can therefore feed back into basic theory and stimulate new laboratory research while increasing the scientific and social relevance of fairness research.

The shift in emphasis is motivated by a simple observation. Fairness concerns change incentives in markets, politics and organizations, thereby influencing aggregate outcomes even when only a subset of individuals care about fairness[4]. When workers reduce effort after perceived unfair pay cuts[5,6], when voters support redistribution conditional on beliefs about merit and opportunity[7–9], when citizens punish free-riders in collective-action settings[10], when young men decide whether to volunteer to join the army to defend their country[11] or when people collectively protest against perceived injustice[12–15], fairness

[1]Department of Economics, University of Zurich, Zurich, Switzerland. [2]Warwick Business School, University of Warwick, Coventry, UK. [3]These authors contributed equally: Ernst Fehr, Keyu Wu. e-mail: ernst.fehr@econ.uzh.ch

https://doi.org/10.1038/s41562-026-02557-7

BOX 1

Glossary of key concepts

Term	Definition
Institution	A stable system of formal rules (for example, laws, contracts and organizations) and informal rules (for example, norms and customs) that structures social interactions, incentives and expectations. Institutions influence which comparisons people make and which fairness ideals become salient.
Social preferences	Preferences in which individuals care not only about their own outcomes but also about the outcomes, intentions or behaviour of others, including concerns for fairness, reciprocity, altruism, guilt and social image.
Reference group (reference agent)	The individual or group against whom people compare their own outcomes when evaluating fairness (for example, coworkers, neighbours, classmates or rich and poor groups in society).
Reference outcome	The specific outcome that an individual regards as fair in a particular situation after applying one or more fairness ideals to the relevant facts and comparison group. Reference outcomes therefore depend on both the fairness ideals individuals hold and the institutional context in which they are applied.
Fairness ideal	A normative principle that specifies how outcomes ought to be evaluated or allocated (for example, equality, merit, need or entitlement). Fairness ideals provide the criteria for determining which outcomes are regarded as fair in a given context.
Egalitarianism	The fairness ideal that equal or more equal outcomes are intrinsically desirable, independent of differences in effort, productivity or initial endowments.
Libertarianism	The fairness ideal that outcomes are fair if they result from voluntary exchange and respect for property rights, regardless of resulting inequality.
Meritocratic ideal	The fairness ideal that rewards should reflect individual effort, contribution, productivity or performance rather than luck or arbitrary circumstances.
Procedural fairness	The perceived fairness of the process by which decisions are made, including whether procedures are consistent, impartial, transparent and respectful and provide opportunities for voice.
Procedural legitimacy	The extent to which people regard decisions or institutional outcomes as acceptable because they were reached through procedures perceived as fair, even when the outcomes themselves are unfavourable.
Norm internalization	The process through which externally enforced social norms become internal motivations, so that individuals continue to comply even when external monitoring or sanctions are absent.
Competitive wage level	The wage predicted by a competitive labour market in which wages adjust to equate labour supply and demand, without fairness concerns or other behavioural frictions.
Welfare weights	The relative normative importance assigned to improving the well-being of individuals at different positions in the income distribution. Empirically measuring these weights provides a way of connecting citizens' distributive preferences to the evaluation of redistributive policies.

concerns together with other social preferences shape wage rigidity, contract performance, tax-and-transfer systems, the success of collective actions, political conflict and political outcomes.

But the very same examples reveal limitations in the current knowledge base. Applying fairness research to real-world institutions requires knowing not only who the relevant comparison targets are—coworkers, neighbours, occupational peers or national groups—but also which reference outcomes people regard as fair; which fairness ideals are activated; how much behavioural weight these ideals receive when self-interest is at stake; which procedures people view as legitimate; how institutions, norms and life experiences shape these motives over time; and which role fairness concerns have when we move beyond dyads and small groups to entire income distributions, where citizens may attend selectively to the poor, the rich or local comparison groups. Without these building blocks, even well-established fairness mechanisms can be difficult to transport from stylized laboratory settings to the explanation of real-world institutions and interactions and the design of policies.

In the remainder of the Perspective, we proceed in three steps. We first summarize established evidence on the economic and political consequences of fairness and other social preferences. We then identify unresolved problems that limit the use of existing insights in explaining real-world events and policy design. Finally, we outline a research agenda for making fairness and equity research more predictive, policy-relevant and useful for designing institutions that are both effective and legitimate.

Established evidence

Below we discuss the role of social preferences in selected economic environments. Interested readers may consult Fehr and Charness[2] for further applications. Throughout, we treat 'fairness and equity preferences' as part of a broader set of social preferences that includes reciprocity, norm enforcement motives and moral emotions such as guilt aversion that discipline behaviour even absent formal sanctions.

Labour markets, wage setting and organizational behaviour

Fairness concerns shape labour market outcomes through reference-dependent perceptions of wages and effort. A classic mechanism is the 'fair wage–effort' hypothesis, which posits that workers compare actual wages to reference wages they perceive as fair and reduce effort when they are paid less. Firms therefore have an incentive to pay wages above the competitive level, providing a behavioural explanation for wage rigidity and unemployment[5,16,17]. Qualitative evidence from firms indicates that managers deliberately avoid nominal wage cuts and strongly unequal pay adjustments because they anticipate negative morale and effort responses rooted in fairness perceptions[5].

A substantial subsequent literature—spanning laboratory experiments, field experiments and administrative data—shows that perceived unfairness has robust behavioural consequences in labour markets. Workers who experience wage policies or pay changes they view as unfair are more likely to quit, withdraw cooperation or engage in counterproductive behaviour, even when such responses are privately costly[6,18–20]. Using large-scale administrative data, Dube et al.[19] show, for

 https://doi.org/10.1038/s41562-026-02557-7

example, that unequal raises within firms substantially increase quit rates among negatively affected workers, highlighting that relative pay comparisons—rather than absolute wage levels—are central for labour supply responses.

Related evidence shows that perceived unfair treatment can reduce productivity and cooperation within teams, particularly when workers observe peers' pay or output and view resulting disparities as unjustified[21,22]. Breza et al.[23] demonstrate, for example, that experimentally induced pay inequality among comparable workers leads to substantial declines in effort and cooperation due to fairness concerns. Overall, perceived unfairness can disrupt workplace cooperation and performance, helping to explain why firms often avoid wage cuts.

Taken together, fairness considerations are a central feature of labour markets. They shape wage-setting practices and influence turnover, productivity and organizational stability. Crucially, these effects arise because workers evaluate wages relative to salient reference standards—coworkers' pay, past pay or perceived entitlements—underscoring the importance of understanding how reference agents and reference outcomes are determined within firms and labour market institutions.

Contracts and institutions as reference-point generators

A striking insight in modern contract theory is that contracts can function as reference points: parties evaluate outcomes not only by material pay-offs but also relative to what they feel entitled to under the agreement[24]. When outcomes fall short of perceived entitlements, aggrievement can lead to shading, sabotage or reduced performance.

Experimental evidence shows that contractual frames can shift perceptions of fairness and, in turn, behaviour—even when the material incentives are unchanged—because contracts influence what parties consider legitimate claims and acceptable deviations[25]. The practical consequence is that legal and organizational design must consider not only incentive compatibility but also fairness compatibility: which agreements generate stable and mutually acceptable reference points?

Yet we still lack a robust formal theory of how individuals select and update reference points in complex environments. This gap matters because many disputes in real labour markets—such as wage negotiations, bonus allocations, lay-offs and promotions—are not simply about resources; they are about contested entitlements.

Collective action, punishment and norm enforcement

Fairness also has an important role in collective action. In public goods settings, many individuals are willing to sanction free-riders, thereby sustaining cooperation far above levels predicted by self-interest alone[10]. Importantly, third-party punishment connects social preferences to social norms: people punish not only to alter immediate incentives but also to enforce perceived normative standards[26]. This norm-enforcement capacity can scale to real-world governance problems—from commons management to civic compliance.

Evidence from field contexts shows how institutional arrangements interact with cooperative norms and punishment to support collective resource management[27–29]. The policy implication is not that punishment is always desirable but that institutional design can harness norm-enforcement motives to reduce enforcement costs and increase compliance—provided the relevant reference standards are legitimate and widely shared.

Redistribution, fairness ideals and political conflict

Distributive politics is perhaps the domain where fairness research could have its largest long-run impact because almost all public policies have distributional implications. Beyond income, wealth taxation and transfer payments, healthcare, climate policy, and access to the legal system all raise distributive questions. Individuals' preferences for redistribution depend not only on self-interest but also on beliefs about merit, luck, opportunity, need and deservingness. Standard political-economy models of redistribution often rely heavily on material self-interest; yet citizens' social preferences and fairness beliefs can shift policy demand and voting behaviour even at constant income.

A recent working paper indicates that fairness ideals differ markedly across contexts and cultures[30], and that the relative salience of merit versus luck can change the acceptance of inequality[31]. This heterogeneity helps to explain why similar objective inequality can yield different social conflicts across societies. Moreover, new work is beginning to connect measured inequality aversion to support for public and private redistribution, strengthening the bridge between individual preferences and collective political demands[9,32,33].

A limitation of much of this experimental literature is that it measures distributive preferences in dyads, while real redistribution is inherently about entire distributions. Recent studies move towards eliciting social welfare weights and distributional preferences over many income classes. These designs reveal that individuals often focus disproportionately on the very poor, the very rich and 'local' neighbours around their own position—suggesting a psychologically grounded, reference-dependent mapping from distributions to policy demand[34]. A recent working paper estimates social marginal welfare weights in representative samples, allowing comparisons with the implicit weights of actual tax systems[35]. Integrating these approaches with modern optimal-tax tools may offer a concrete path towards a welfare economics that respects citizens' social preferences rather than assuming people only care about their own benefits[36,37].

The main findings reviewed in this section and the broader literature on social preferences are summarized in Box 2.

Unresolved problems

Moving from documenting fairness concerns to applying them requires solving five foundational problems that remain empirically largely unresolved. First, fairness judgements presuppose reference agents—yet who people compare themselves to varies by institutional setting and the social environment. Second, fairness judgements depend on fair reference outcomes, the behavioural weight placed on them and the perceived legitimacy of the procedure that produced them, all of which may shift with context. Third, fairness motives may be shaped through childhood and institutional experience, yet their developmental origins are not well understood. Fourth, the relationship between norms and preferences remains opaque; in particular, we know very little about when and under which conditions norms are internalized such that they become features of individuals' social preferences. Finally, distributive judgements in society concern entire income distributions, but the methods we currently use are often not yet scaled to that task. Together, these questions constitute the conceptual bottlenecks that must be addressed to productively incorporate fairness into institutional and policy design.

BOX 2

What we already know about social preferences and fairness

- Social preferences are genuine motives. In anonymous one-shot experiments, many people give to others, reject unfair offers, reciprocate trust, contribute to public goods and punish unfair behaviour, even when doing so is costly and cannot improve future reputation or material outcomes[1,2].
- Social preferences are widespread. Many people care not only about their own material outcomes but also about the outcomes of others. Broad population evidence indicates that purely self-interested individuals are a minority, while individuals with social preferences are common[2,30,32].
- Theory predicts that institutions shape the consequences of social preferences. Stable social preferences can generate very different behaviours across strategic environments. For example, competition may make fair-minded individuals behave as if they were self-interested, whereas fair-minded individuals can sustain cooperation when peer punishment opportunities exist[4,63].
- Social preferences support cooperation. Altruism can directly promote cooperation, while inequality aversion and negative reciprocity can sustain cooperation by motivating punishment of free-riders. Peer pressure and peer punishment can support large-scale collective action[4,10,64,65].
- Fairness concerns shape labour markets. Workers compare their wages with those of relevant peers and react negatively to wage inequality that appears unjustified. Disadvantageous wage inequality can reduce effort, job satisfaction, social cohesion and cooperation, and can increase quitting[6,18–21,23].
- Resistance to wage cuts helps to explain wage rigidity. Workers often view nominal wage cuts as unfair and may respond with lower effort, sabotage, theft or reduced product quality. Firms may therefore avoid wage cuts even when standard competitive models predict wage flexibility[5,16,17,66].
- Social preferences affect who gets hired and rewarded. When employers can observe workers' cooperative or reciprocal tendencies, they may offer more attractive jobs to workers who have shown high effort or reciprocity. Positively reciprocal workers can receive higher wages, more rent sharing and greater discretion[67–69].
- Social preferences matter for contracts and incentives. Explicit incentives and control can sometimes undermine voluntary cooperation, whereas trust, informal bonuses and informal sanctions can improve performance. Social preferences help to explain incomplete contracts, fixed wages and informal rewards, and explicit contracts may themselves constitute a reference standard[24,25,70,71].
- Social preferences influence politics and collective conflict. Non-selfish motives such as altruism, fairness concerns and other moral commitments motivate costly participation in political movements and protests[12,14,15]. Prosocial motives can therefore contribute to the provision of democratic rights and other political public goods.
- Social preferences shape support for redistribution. Inequality aversion and altruism predict greater support for redistributive policies, especially among higher-income individuals for whom redistribution is materially costly. Meritocratic beliefs can reduce such support when income differences are seen as deserved[9,32,33].

Box 3 summarizes these unresolved problems and illustrates the kinds of questions that require more in-depth research.

Identifying the relevant reference agents

Distributional preference models typically assume that individuals compare outcomes to a set of 'relevant others'. In experiments, the relevant others are often simply assumed to be the participants' interaction partners. But in reality, reference groups are endogenous and context-dependent: neighbours can become salient for housing, safety and local status; coworkers for wages and promotions; classmates for grades; and broader national groups for redistribution debates. Economic and institutional context—segregation, workplace structure, transparency policies, social media exposure and narratives—is very likely to shape which comparisons are made and which are emotionally charged. Moreover, economic and political actors may deliberately shape who the relevant others are.

Recent meta-analyses on social-comparison research suggests that reference agents are unlikely to be drawn from the full population. Comparisons should be strongest when targets are similar, local, socially proximate and informationally diagnostic, and when institutions make previously opaque outcomes visible[23,38,39]. This is confirmed by a recently published paper[40] documenting that people view inequalities within their coworkers and education groups as considerably more unfair than overall inequality. A recent working paper[41] found substantial variation in the effects of rank information on satisfaction across reference groups, with national reference groups showing the weakest effects. These findings also highlight a major challenge: if researchers or policymakers mis-specify which reference groups are relevant, they risk misinterpreting fairness perceptions and predicting behavioural responses as well as welfare implications incorrectly.

Identifying reference outcomes and operative fairness ideals

Even if reference agents are known, fairness evaluations depend on reference outcomes: what would have been fair (equal splits, proportional-to-contribution rules, need-based allocations or entitlement-based claims). The empirical identification of reference outcomes is particularly difficult because outcomes are often multidimensional and observed actions often reflect a mixture of motives—distributional concerns, reciprocity, efficiency, image and strategic incentives.

A prominent approach in the fairness-ideals literature assumes exogenously given ideals—for example, libertarian, egalitarian or meritocratic—and then estimates the population shares of these types from impartial-spectator allocations or related designs[42,43]. This framework has been enormously useful, including in documenting cross-cultural variation and the role of merit versus luck in inequality acceptance. But the exogeneity assumption is also a conceptual bottleneck.

A central unresolved issue is the joint identification of (1) which fairness ideals individuals hold and (2) the behavioural weight these ideals receive when individuals' material self-interest is at stake. Most existing designs either isolate normative judgements by removing self-interest or study behaviour under self-interest without being able to disentangle the underlying fairness ideal from its motivational force. As a result, observed deviations from fairness in stakeholder settings may reflect weaker commitment to a given ideal, the activation of a different ideal or strategic responses to social or material incentives.

https://doi.org/10.1038/s41562-026-02557-7

BOX 3

Examples of unresolved problems

- Who are the relevant reference agents? Fairness judgements require comparison targets, but in real life these targets are often unknown and endogenous, contingent on context. Workers may compare themselves to teammates, occupational peers, past selves or workers in other firms; citizens may compare themselves to neighbours, educational peers or the national income distribution. As an example, when a firm introduces pay transparency, will workers react most strongly to wage gaps within their team, within their occupation or across the whole firm?
- What are the relevant reference outcomes? Even if the comparison group is known, it is often unclear what outcome people regard as fair. They may apply equality, proportionality to effort, compensation for need, equal opportunity, entitlement-based claims or mixtures of these principles. As an example, when workers oppose unequal bonuses, do they object because bonuses violate equality, because measured performance is viewed as inaccurate or because past promises created different entitlements?
- How do fairness ideals and self-interest jointly shape behaviour across contexts? Many experiments identify fairness ideals in spectator settings, where individuals judge allocations between others without their own outcomes being involved. But in real life people are often stakeholders with material interests. Observed behaviour may therefore reflect the fairness ideal people hold, the weight they attach to it or strategic incentives that incorporate self-interest. As an example, if a person endorses the meritocratic allocation as a spectator but rejects it when personally involved, did the fairness ideal change or did the weight attached to merit now decline because they also want to maximize their own benefit?
- How do procedures affect the legitimacy of outcomes? Institutional outcomes are produced through procedures, and people evaluate voice, neutrality, consistency, respectful treatment and explanations as well as final allocations. Yet we know too little about how procedural and distributive fairness interact. As an example, when do fair procedures make lay-offs, tax increases or welfare cuts more acceptable, and when are unfavourable outcomes so severe that procedural fairness no longer preserves legitimacy?
- How are fairness and social preferences formed? Developmental evidence shows that children acquire fairness concepts and prosocial tendencies early[47,48], but it remains unclear how families, peers, schools, firms and historical institutions shape the content, strength and persistence of social preferences. As an example, does cooperative learning, exposure to diversity or teacher enforcement of rules durably affect altruism, reciprocity or meritocratic fairness ideals into adulthood?
- When do social norms become internalized preferences? Norms can constrain behaviour through sanctions and expectations, while preferences can motivate norm compliance even without external enforcement. The causal relationship between the two remains poorly understood. As an example, when a school, firm or community repeatedly enforces a fairness norm, does this merely change behaviour within that environment, or does it create a durable preference that persists across new settings, new interaction partners and changing external incentives?
- How do people evaluate whole income distributions? Real redistribution concerns large populations, not only dyads or small groups. People may simplify complex distributions by focusing on the very poor, the very rich or local neighbours near their own position, but we lack a general model of these heuristics. As an example, do citizens support redistribution because they dislike overall inequality, because they care about poverty, because they resent top incomes or because policies or political narratives affect salient local comparison groups?

More broadly, fairness ideals may not be exogenously given. In many real-world settings, individuals may be influenced by multiple ideals whose salience depends on framing, group identity and institutional context. Fairness ideals should thus be inferred across tasks and contexts rather than assumed to be fixed types. In addition, fairness ideals should be understood more broadly than distributive rules such as equality, equity, meritocracy or need. Organizational justice research[44–46] shows that, in many institutional settings, people also evaluate the procedures through which outcomes are produced. Voice, consistency, neutrality, respectful treatment and adequate explanations can determine whether even unfavourable outcomes are viewed as legitimate or unfair, thus indicating that studying the interaction between procedural and distributive fairness is an important direction for future research.

How fairness and social preferences are formed

If social preferences shape major societal outcomes, their developmental origins become first-order questions. Research in developmental psychology[47,48] shows that children acquire fairness concepts and prosocial tendencies early, but fairness knowledge and fair behaviour do not always develop in parallel. Work in social-domain theory[49–51] suggests that schools shape fairness development not only by rewarding or punishing behaviour but also by teaching children how to distinguish moral concerns about harm, rights and fairness from social conventions. Cross-cultural evidence[52] suggests that large-scale institutions, such as markets, religions and community structure, are associated with variation in fairness and punishment towards strangers, reinforcing the idea that social preferences are partly shaped by institutional environments rather than fixed individual traits.

Further evidence indicates that role models and education in early childhood can have measurable causal effects on social preferences, suggesting that institutions may causally shape other-regarding motivations rather than merely sorting individuals into environments[53,54]. Yet we lack causal evidence about which developmental inputs matter most (parental modelling, peer interactions, formal rules, punishment systems or cultural narratives) and how they matter, how persistent these effects are, and how they interact with later experiences in labour markets, organizations or politics.

Social preferences may also remain malleable beyond childhood. For example, organizational settings may influence social preferences through the structure of rewards, incentives and workplace culture, raising the question of whether firms and other organizations can actively foster or erode prosocial motivations. Similarly, major life events and shocks—such as adverse health or income shocks, family disruption or changes in community attachment—may alter trust, reciprocity and fairness concerns, yet we know very little about these dynamics.

There is also emerging evidence that social interaction across social, economic or caste boundaries can shape fairness preferences. Exposure to disadvantaged peers, for instance, has been shown to increase prosociality and reduce discriminatory behaviour among more advantaged individuals, suggesting that social integration may have a key role in the formation of social preferences[55].

Beyond individual experiences, a society's economic and governance institutions may exert powerful long-run effects on social preferences. Recent work suggests that market participation and economic integration can foster prosocial motivations, while histories of self-governance and democratic interaction tend to promote conditional cooperation and other forms of prosociality[56,57]. These findings are consistent with broader historical evidence indicating that past institutional experiences—such as periods of political autonomy—are associated with persistent differences in prosocial behaviour at the community level.

Taken together, these observations suggest that fairness and social preferences are not fixed traits but are shaped by early-life experience, social interactions, institutional environments and historical experience. Understanding which forces matter most, how persistent their effects are and through which psychological mechanisms they operate remains a major unresolved challenge—and a prerequisite for assessing whether and how policy and institutional design can durably influence social preferences.

Social preferences and social norms

Social norms are commonly known standards of behaviour that are based on widely shared views of how individual group members ought to behave in a given situation. They are thus properties of groups, whereas social preferences are properties of individuals. Yet they are deeply intertwined. Norms constrain behaviour through social approval, sanctions and expectations; preferences motivate behaviour through intrinsic valuations, including a willingness to punish norm violations or to feel guilt when disappointing others.

A central unresolved question is whether persistent norms, acting as durable constraints, eventually shape preferences via motivational internalization. Conversely, do social preferences help to enforce norms? Work on (third-party) punishment already suggests a tight link between sanctioning behaviour and norm enforcement[10,26]. Relatedly, guilt aversion[58] and other forms of social preferences[59] support norm compliance: individuals may keep promises or behave fairly because violating norms triggers aversive feelings and thus compliance even without external punishment. These ideas are powerful but under-integrated: we still lack clean empirical evidence on the psychological processes and causal drivers of norm internalization.

Distributional preferences for whole income distributions

Perhaps the most under-researched topic—relative to its policy importance—is how individuals evaluate the distribution of income in large groups. Classic dictator, ultimatum and trust games focus on pairs, and even public goods experiments typically involve small groups. By contrast, real-world redistribution concerns entire income distributions, spanning many ranks and dimensions, including the treatment of very poor or very rich groups, social mobility and opportunity.

Assessing whole income distributions is, however, a task of daunting cognitive complexity. Individuals must evaluate outcomes for many agents simultaneously, often under uncertainty about effort, luck and opportunity. It is therefore unlikely that people process full distributions in a comprehensive or welfare-theoretic manner. Instead, individuals may simplify the task by focusing on a small set of salient comparison objects. In view of the complexity of this task, it appears plausible that the very bottom and the very top of the income distribution, as well as individuals' immediate neighbours around their own position, serve as focal reference agents for assessing societal fairness.

Recent experimental evidence provides initial support for this view. Using designs that allow participants to reallocate income across many ranks, Fisman et al.[60] show that distributional preferences are disproportionately shaped by concern for the poorest and richest groups and local comparisons near one's own income position, rather than by sensitivity to the entire distribution or to summary inequality measures. These findings suggest that what is often interpreted as concern for aggregate inequality may instead reflect reference-based evaluations anchored on salient distributional positions.

This interpretation has important implications for both theory and policy. It implies that citizens' responses to inequality and redistribution may depend less on abstract measures of inequality and more on how policies affect salient reference groups. At the same time, it raises unresolved questions. How stable are such salience-based reference agents across societies and institutional environments? To what extent do information policies or political framing shift attention towards particular distributional positions? How do beliefs about mobility and opportunity interact with attention to the top, bottom or local neighbours in shaping redistributive preferences?

Addressing these questions requires moving beyond pairwise allocation tasks towards experimental and field designs that explicitly manipulate informational complexity, salience and framing in large-group settings.

A forward-looking research agenda

The unresolved problems discussed above are tightly connected and point towards a common challenge: fairness and equity judgements in real-world environments are shaped under cognitive, informational, social and institutional constraints. The central questions thus become: which reference points do institutions create, which fairness ideals do they activate, which procedures are perceived as legitimate and how do fairness, norms and preference formation interact to shape behaviour in labour markets, organizations and redistributive politics? Addressing these questions requires integrating experiments with administrative data, institutional variation and structural approaches capable of identifying, for example, latent reference points, operative fairness ideals and procedural legitimacy. A forward-looking agenda must therefore study reference groups, reference outcomes and fairness ideals as selected, updated and stabilized by institutions over time (Table 1).

Identifying reference agents in the field

A first priority is to develop new formal models where individuals endogenously choose their reference groups and and analyse how this choice affects their fairness perceptions. A second priority is empirical: identify reference groups by combining experimental elicitation with observational data. Several strategies are promising:

1. Conceptual and theoretical advances. Models that include the choice or activation of reference agents within institutions can clarify the mechanisms that determine who counts as a comparison target in practice.
2. Institutional or informational shocks that shift salience. For example, transparency, team reorganization or neighbourhood change alters who is compared to whom. Estimating behavioural responses to such shocks can reveal the reference set.
3. Network-based designs. Administrative and digital trace data can map social proximity (such as coworker networks or neighbourhood contacts) and test whether comparisons track network structure rather than demographic similarity alone.
4. Multi-context measurement within individuals. Repeated elicitation across candidate reference groups—coworkers versus neighbours versus national distribution—can be combined with latent-variable models to estimate individual-specific reference weights.

Once reference agents are identified, fairness predictions become sharper, and policy design can target the information and institutional levers that govern comparison sets.

Table 1 | A research agenda for the next generation of fairness research

Research objective	Key research directions
Identify reference agents in the field	Explain how institutions select, activate and stabilize the comparison groups that shape fairness judgements in labour markets, neighbourhoods, organizations and redistributive politics. Develop theoretical models of endogenous reference agents, exploit institutional and informational shocks (for example, pay transparency, team reorganization and neighbourhood change) and combine network-based measures with multi-context elicitation of individual-specific reference weights.
Identify fairness ideals, motivational weights and procedural legitimacy	Distinguish which fairness ideals individuals apply without exogenous assumptions on the set of ideals, how much behavioural weight these ideals receive when self-interest is at stake and how procedures affect the legitimacy of outcomes. Link impartial-spectator judgements to stakeholder choices, jointly estimate fairness ideals and motivational weights, and examine how voice, neutrality, consistency, respectful treatment and explanations shape acceptance of otherwise identical outcomes.
Study institutions as preference-formation environments	Treat schools, firms, communities and early-childhood programmes not only as settings in which social preferences are expressed but also as environments in which they are formed, strengthened or weakened. Use longitudinal, field-experimental and quasi-experimental designs to identify the mechanisms of preference formation, assess the persistence of these effects and examine whether common institutional rules shape fairness ideals differently across socio-economic backgrounds.
Trace the internalization and enforcement of social norms	Clarify when social norms merely constrain behaviour through expectations, observability and sanctions, and when they become internalized as social preferences. Combine measures of perceived norms, sanctioning expectations and internal motivations with variation in observability and enforcement regimes to identify when norm-compliant behaviour persists after external enforcement is removed.
Scale fairness research to whole income distributions and policy choice	Move beyond dyads and small groups towards the study of society-wide income distributions and redistributive politics. Develop models with society-wide reference agents, estimate distributive weights across countries, integrate beliefs about mobility and opportunity, and link elicited distributional preferences to support for taxes, transfers and other redistributive policies.

Identifying fairness ideals and procedural legitimacy

As mentioned above, a central agenda is to identify both the distributive ideals people apply and the procedural standards through which outcomes are judged as legitimate. To this end, we suggest the following strategies:

- Within-individual designs linking spectator and stakeholder roles. Experimental designs that elicit fairness judgements as an impartial spectator and then observe behaviour in stakeholder settings can help to disentangle which fairness ideal is applied from how much behavioural weight it receives when self-interest is at stake.
- Joint estimation of fairness ideals and motivational weights. Rather than treating fairness ideals as fixed types, future work could jointly infer the operative fairness principle and its behavioural weight from observed choices, allowing both to vary across contexts and decision roles.
- Experimental variation in procedures as well as outcomes. Future studies should vary not only the allocation itself but also the procedure that produces it. This would show when voice, consistency, neutrality, respectful treatment or explanations mitigate resistance to unfavourable outcomes, and when unfavourable outcomes are too severe for fair procedures to sustain legitimacy.
- Multi-context elicitation to study ideal activation. Repeated measurement across settings that vary in self-interest, strategic interaction or institutional framing can reveal when individuals apply the same fairness ideal across roles and when different ideals become salient.

Developmental and institutional determinants of preference formation

A major opportunity is to treat schools and early-childhood programmes as environments where social preferences are formed and where causal identification may be possible[53]. Evidence that early-childhood education affects social preferences could be expanded along three dimensions:

- Mechanisms: is it cooperative play, teacher enforcement of rules, exposure to diversity or structured norm-sharing that drive change?
- Dynamics: are effects persistent into adulthood, or do later environments overwrite early preferences?
- Interaction with inequality: do children from different socio-economic backgrounds form different fairness ideals under the same institutional rules?

Connecting norms and preferences by tracing internalization and enforcement

A key task is to determine whether social norms merely constrain behaviour through external rewards and sanctions or whether persistent exposure to norms gradually transforms them into internal motivations. Understanding this process of motivational internalization is essential for explaining both the formation of social preferences and the long-run effectiveness of institutions.

To connect social norms and social preferences, we need longitudinal designs that jointly measure norms, behaviour and motivation. A feasible research programme could consider the following strategies:

- Belief elicitation about norms and expectations. Measure whether changes in individuals' perceptions of prevailing norms translate into subsequent changes in their own social preferences.
- Behaviour under varying observability and sanction regimes. Study whether cooperative and fair behaviour persists once external monitoring or sanctions are removed, providing evidence on whether norms have become internalized.
- Dynamic exposure to stable norms. Examine whether prolonged exposure to environments with stable cooperative norms (for example, in schools, workplaces or communities with strong cooperative traditions) strengthens social preferences and identify which institutional features—peer interactions, leadership, democratic participation or sanctioning practices—are most effective in fostering lasting internalization.

Together, these research strategies would make it possible to distinguish behaviour that is sustained primarily by external rewards and sanctions from behaviour that is driven by internalized social preferences. They would also help to identify the psychological mechanisms through which norms become motivationally effective. Guilt aversion and negative reciprocity provide particularly promising bridge concepts because they link socially shared expectations to internal emotional responses. Once social norms have been internalized, violating them may trigger guilt, while observing violations by others may evoke anger and a willingness to punish, thus contributing to prosocial cooperation even in the absence of external enforcement[10,58]. Understanding

 https://doi.org/10.1038/s41562-026-02557-7

how these motivational mechanisms emerge and evolve would help to unify currently separate literatures on social norms, social preferences and moral emotions into a common framework for explaining long-run cooperation and institutional effectiveness.

Applications in welfare economics and political economics

A substantial application gap also exists in political and welfare economics. If citizens care about inequality, poverty and relative position, then normative analysis that assumes purely self-interested utility misses morally relevant objects of choice. Calls to bring social concerns into public economics and political economics have become increasingly explicit[9,32,33,36,37], including recent working papers that examine the consequences of other-regarding preferences for assessing overall welfare[61,62].

Recent work on society-wide reference agents[40,41] and distributive preferences[60] offers the prospect of grounding redistributive policy more firmly in empirically measured fairness preferences. A promising step is to estimate the welfare weights citizens attach to improving the situation of people at different positions in the income distribution, rather than assuming these weights in advance[35–37]. This requires measurement beyond two-person allocation games. A recent working paper estimating such weights[35] in representative samples and studies documenting how attention focuses on salient positions[60] in the income distribution represent important first steps.

The next steps could include:

- Developing theoretical models where individuals focus on salient positions in the income distribution.
- Identifying the empirical distribution of welfare weights across countries.
- Identifying how the welfare weights change with perceptions about equality of opportunity and intergenerational mobility.
- Linking elicited weights to support for concrete policies such as income taxes or inheritance taxation.
- Studying how reference groups shape preferences for redistribution.

A complementary direction is political economics: redistribution is implemented through voting and party competition, not a benevolent planner. Social preferences can shape coalition formation, polarization and the stability of welfare states. Bringing distributional-preference measurement into models of political choice—while accounting for endogenous beliefs and narratives about merit—should be a central agenda item.

The next stage of fairness research

The fairness literature has reached a stage where progress increasingly depends on integration and application. The central open questions no longer concern whether people care about fairness, but how fairness judgements are selected, shaped and stabilized across contexts. Who becomes a relevant reference agent, which outcomes are taken as fair, which fairness ideals are activated, how social preferences are formed and how individual motivations interact with social norms all depend on institutional, informational and social environments.

A useful vision is to treat fairness as a core component of how humans navigate social life—and to treat institutions as fairness technologies that select reference points, activate norms and shape the incentives and meanings of distributive choices. This view does not imply that policy should always follow majority fairness intuitions; fairness ideals can conflict, and some may be ethically problematic. But it does imply that economists and behavioural scientists should stop treating fairness as peripheral when analysing labour markets, organizations and redistribution.

The benefit is practical. Better knowledge about reference groups and reference outcomes can improve wage policies and transparency rules. Better knowledge about the formation of social preferences can inform education and integration policy. Better integration of distributional preferences into welfare economics can generate policy analyses that are descriptively realistic and normatively responsive to what citizens value. If these frontiers are pursued, fairness research can inform the design of policies that are both effective and perceived as legitimate.

References

1. Kagel, J. H. & Cooper, D. J. in *The Handbook of Experimental Economics*, *Volume 2* (eds Kagel, J. H. & Roth, A. E.) 217–289 (Princeton Univ. Press, 2017).
2. Fehr, E. & Charness, G. Social preferences: fundamental characteristics and economic consequences. *J. Econ. Lit.* **63**, 440–514 (2025).
3. Fehr, E. & Krajbich, I. in *Neuroeconomics—Decision Making and the Brain* (eds Glimcher, P. W. & Fehr, E.) 193–218 (Academic Press, 2014).
4. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
5. Bewley, T. F. *Why Wages Don't Fall During a Recession* (Harvard Univ. Press, 1999).
6. Kube, S., Marechal, M. A. & Puppe, C. Do wage cuts damage work morale? Evidence from a natural field experiment. *J. Eur. Econ. Assoc.* **11**, 853–870 (2013).
7. Alesina, A. & Angeletos, G. M. Fairness and redistribution. *Am. Econ. Rev.* **95**, 960–980 (2005).
8. Almas, I., Cappelen, A. W. & Tungodden, B. Cutthroat capitalism versus cuddly socialism: are Americans more meritocratic and efficiency-seeking than Scandinavians? *J. Polit. Econ.* **128**, 1753–1788 (2020).
9. Fehr, E., Epper, T. & Senn, J. Social preferences and redistributive politics. *Rev. Econ. Stat.* https://doi.org/10.1162/rest_a_01511 (2024).
10. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
11. Jiang, G. *No Fairness, No Cooperation: Draft Dodging by the Rich and Voluntary Enlistment in World War II*. Doctoral dissertation, Univ. Zurich (2024).
12. Opp, K. D. *Theories of Political Protests and Social Movements* (Routledge, 2009).
13. Finkel, S. E., Muller, E. N. & Opp, K. D. Personal influence, collective rationality, and mass political-action. *Am. Polit. Sci. Rev.* **83**, 885–903 (1989).
14. Passarelli, F. & Tabellini, G. Emotions and political unrest. *J. Polit. Econ.* **125**, 903–946 (2017).
15. Cantoni, D., Heizlsperger, L. J., Yang, D. Y., Yuchtman, N. & Zhang, Y. J. The fundamental determinants of protest participation: evidence from Hong Kong's antiauthoritarian movement. *J. Public Econ.* **211**, 104667 (2022).
16. Akerlof, G. A. & Yellen, J. L. The fair wage–effort hypothesis and unemployment. *Q. J. Econ.* **105**, 255–283 (1990).
17. Kaur, S. Nominal wage rigidity in village labor markets. *Am. Econ. Rev.* **109**, 3585–3616 (2019).
18. Card, D., Mas, A., Moretti, E. & Saez, E. Inequality at work: the effect of peer salaries on job satisfaction. *Am. Econ. Rev.* **102**, 2981–3003 (2012).
19. Dube, A., Giuliano, L. & Leonard, J. Fairness and frictions: the impact of unequal raises on quit behavior. *Am. Econ. Rev.* **109**, 620–663 (2019).
20. Bandiera, O., Barankay, I. & Rasul, I. Social preferences and the response to incentives: evidence from personnel data. *Q. J. Econ.* **120**, 917–962 (2005).
21. Cohn, A., Fehr, E., Herrmann, B. & Schneider, F. Social comparison and effort provision: evidence from a field experiment. *J. Eur. Econ. Assoc.* **12**, 877–898 (2014).

22. Cullen, Z. & Perez-Truglia, R. How much does your boss make? The effects of salary comparisons. *J. Polit. Econ.* **130**, 766–822 (2022).
23. Breza, E., Kaur, S. & Shamdasani, Y. The morale effects of pay inequality. *Q. J. Econ.* **133**, 611–663 (2018).
24. Hart, O. & Moore, J. Contracts as reference points. *Q. J. Econ.* **123**, 1–48 (2008).
25. Fehr, E., Hart, O. & Zehnder, C. Contracts as reference points—experimental evidence. *Am. Econ. Rev.* **101**, 493–525 (2011).
26. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
27. Ostrom, E. Collective action and the evolution of social norms. *J. Econ. Perspect.* **14**, 137–158 (2000).
28. Rustagi, D., Engel, S. & Kosfeld, M. Conditional cooperation and costly monitoring explain success in forest commons management. *Science* **330**, 961–965 (2010).
29. Kosfeld, M. & Rustagi, D. Leader punishment and cooperation in groups: experimental field evidence from commons management in Ethiopia. *Am. Econ. Rev.* **105**, 747–783 (2015).
30. Almas, I., Cappelen, A. W., Sorensen, E. O. & Tungodden, B. *Fairness Across the World*. Discussion Paper 03/23 (Norwegian School of Economics, 2023).
31. Andre, P. Shallow meritocracy. *Rev. Econ. Stud.* **92**, 772–807 (2025).
32. Kerschbamer, R. & Muller, D. Social preferences and political attitudes: an online experiment on a large heterogeneous sample. *J. Public Econ.* **182**, 104076 (2020).
33. Epper, T. F. et al. Inequality aversion predicts support for public and private redistribution. *Proc. Natl Acad. Sci. USA* **121**, e2401445121 (2024).
34. Charite, J., Fisman, R. & Kuziemko, I. Reference points and redistributive preferences: experimental evidence. *J. Public Econ.* **216**, 104761 (2023).
35. Capozza, F. & Srinivasan, K. *Who Should Get the Money? Estimating Welfare Weights in the U.S.* Discussion Paper No. 50 (Univ. Zurich, Research Priority Program on Equality of Opportunity, 2024).
36. Saez, E. & Stantcheva, S. Generalized social marginal welfare weights for optimal tax theory. *Am. Econ. Rev.* **106**, 24–45 (2016).
37. Saez, E. Public economics and inequality: uncovering our social nature. *AEA Pap. Proc.* **111**, 1–26 (2021).
38. Gerber, J. P., Wheeler, L. & Suls, J. A social comparison theory meta-analysis 60+ years on. *Psychol. Bull.* **144**, 177–197 (2018).
39. Crusius, J., Corcoran, K. & Mussweiler, T. in *Theories of Social Psychology* (ed. Chadee, D.) 165–187 (Wiley-Blackwell, 2022).
40. Hvidberg, K. B., Stantcheva, S. & Thustrup-Kreiner, C. Social position and fairness views on inequality. *Rev. Econ. Stud.* **90**, 3083–3118 (2023).
41. Xu, X. et al. Which income comparisons matter to people, and how? Evidence from a large field experiment. *Am. Econ. J. Appl. Econ.* **18**, 145–177 (2026).
42. Cappelen, A. W., Hole, A. D., Sorensen, E. O. & Tungodden, B. The pluralism of fairness ideals: an experimental approach. *Am. Econ. Rev.* **97**, 818–827 (2007).
43. Cappelen, A. W., Konow, J., Sorensen, E. O. & Tungodden, B. Just luck: an experimental study of risk-taking and fairness. *Am. Econ. Rev.* **103**, 1398–1413 (2013).
44. Brockner, J. & Bobocel, D. R. How justice theory and research can help address organizational societal problems. *Organ. Psychol. Rev.* **14**, 288–306 (2024).
45. Brockner, J. & Wiesenfeld, B. M. in *Social Psychology and Justice* (ed. Lind, E. A.) 213–242 (Routledge, 2020).
46. van den Bos, K. *The Fair Process Effect: Overcoming Distrust, Polarization, and Conspiracy Thinking* (Cambridge Univ. Press, 2024).
47. McAuliffe, K., Blake, P. R., Steinbeis, N. & Warneken, F. The developmental foundations of human fairness. *Nat. Hum. Behav.* **1**, 0042 (2017).
48. Blake, P. R., McAuliffe, K. & Warneken, F. The developmental origins of fairness: the knowledge–behavior gap. *Trends Cogn. Sci.* **18**, 559–561 (2014).
49. Yoo, H. N. & Smetana, J. G. Distinctions between moral and conventional judgments from early to middle childhood: a meta-analysis of social domain theory research. *Dev. Psychol.* **58**, 874–889 (2022).
50. Killen, M., Elenbaas, L. & Rizzo, M. T. Children as agents of change: moral reasoning and group knowledge promote social justice in childhood. *Hum. Dev.* **61**, 281–296 (2018).
51. Killen, M. & Rutland, A. *Children and Social Exclusion* (Wiley-Blackwell, 2011).
52. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
53. Kosse, F., Deckers, T., Pinger, P., Schildberg-Horisch, H. & Falk, A. The formation of prosociality: causal evidence on the role of social environment. *J. Polit. Econ.* **128**, 434–467 (2020).
54. Cappelen, A., List, J., Samek, A. & Tungodden, B. The effect of early-childhood education on social preferences. *J. Polit. Econ.* **128**, 2739–2758 (2020).
55. Rao, G. Familiarity does not breed contempt: generosity, discrimination, and diversity in Delhi schools. *Am. Econ. Rev.* **109**, 774–809 (2019).
56. Enke, B. Market exposure and human morality. *Nat. Hum. Behav.* **7**, 134–141 (2023).
57. Rustagi, D. Historical self-governance and norms of cooperation. *Econometrica* **92**, 1473–1502 (2024).
58. Charness, G. & Dufwenberg, M. Promises and partnership. *Econometrica* **74**, 1579–1601 (2006).
59. Ellingsen, T. & Mohlin, E. A model of social duties. *J. Polit. Econ.* **134**, 1307–1346 (2026).
60. Fisman, R., Kuziemko, I. & Vannutelli, S. Distributional preferences in larger groups: keeping up with the Joneses and keeping track of the tails. *J. Eur. Econ. Assoc.* **19**, 1407–1438 (2021).
61. Aronsson, T. & Johansson-Stenman, O. *Optimal Second-Best Taxation When Individuals Have Social Preferences*. Working Paper (Umea School of Business, Department of Economics, 2020).
62. Eden, M. & Piacquadio, P. G. *The Ethical Mirror*. URPP Equality of Opportunity Discussion Paper Series (URPP, 2024).
63. Falk, A. & Fischbacher, U. A theory of reciprocity. *Game Econ. Behav.* **54**, 293–315 (2006).
64. Gachter, S. & Thoni, C. Social learning and voluntary cooperation among like-minded people. *J. Eur. Econ. Assoc.* **3**, 303–314 (2005).
65. Kölle, F., Quercia, S. & Tripodi, E. Social preferences under the shadow of the future. *Exp. Econ.* **28**, 22–42 (2025).
66. Breza, E., Kaur, S. & Shamdasani, Y. Labor rationing. *Am. Econ. Rev.* **111**, 3184–3224 (2021).
67. Bartling, B., Fehr, E. & Schmidt, K. M. Screening, competition, and job design: economic origins of good jobs. *Am. Econ. Rev.* **102**, 834–864 (2012).
68. Dohmen, T., Falk, A., Huffman, D. & Sunde, U. Homo reciprocans: survey evidence on behavioural outcomes. *Econ. J.* **119**, 592–612 (2009).
69. Schneider, F. H., Brun, F. & Weber, R. A. Sorting and wage premiums in immoral work. *Rev. Econ. Stat.* https://doi.org/10.1162/rest_a_01536 (2024).
70. Benabou, R. & Tirole, J. Incentives and prosocial behavior. *Am. Econ. Rev.* **96**, 1652–1678 (2006).
71. Fehr, E., Gachter, S. & Kirchsteiger, G. Reciprocity as a contract enforcement device: experimental evidence. *Econometrica* **65**, 833–860 (1997).

https://doi.org/10.1038/s41562-026-02557-7

Author contributions
Both authors contributed equally to all aspects of the paper.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence should be addressed to Ernst Fehr.

Peer review information *Nature Human Behaviour* thanks Jessica Sommerville and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.