Ideology Critique and Game Theory

Author: Jacob Barrett

Email: jacob.barrett@philosophy.ox.ac.uk

Affiliation: University of Oxford

Abstract: Ideology critics believe that many bad social practices persist because of ideology, and that critiquing ideology is an effective way to promote social reform. Skeptics draw on game theory to argue that the persistence of such practices is better explained by collective action problems, and that ideology critique is causally inefficacious. In this paper, I reconcile these camps. I show that while game theory can help us to identify contexts where ideology critique makes no difference, it also reveals causal mechanisms by which ideology critique can have a significant effect. I then consider some objections and directions for further research.

Key words: ideology critique, ideology, epistemic distortion, game theory, collective action

1. Introduction

Ideology critique is experiencing a renaissance in analytic philosophy. Spurred largely by the efforts of thinkers such as Sally Haslanger (2012; 2017) and Tommie Shelby (2003; 2016), more and more philosophers have begun to turn their attention to criticizing ideologies—that is, to attempting to expose the collectively held false or irrational beliefs and other attitudes that distort our understanding of the social world.[1] These ideology critics see themselves as

[1] See also Stanley (2015), Srinivasan (2016) and, for other recent work falling less squarely in the analytic tradition, Celikates (2006) and Jaeggi (2008). The term "ideology" is highly contested, but I follow all such thinkers in understanding ideology as a collective epistemic distortion. I therefore use the term "in the pejorative sense" rather than, more neutrally, as a

engaged in an ameliorative project, because they believe that ideologies often serve to stabilize unjust or otherwise suboptimal arrangements, and that criticizing ideologies can therefore help to destabilize these arrangements. If people can only see their way through ideological distortion, the thought goes, then things will change for the better.

Not everyone is so taken with this project. Drawing on tools from game theory, Joseph Heath (2000; 2001), Michael Rosen (1996, ch. 8), and others have argued that ideology critique is an ineffective way to promote social reform. In many cases where we may be tempted to think that ideology explains the persistence of bad arrangements, a better explanation is that individuals face a collective action problem. More recently, Kirun Sankaran (2020) has gone so far as to claim that, even in cases where ideology is at play, merely exposing it has little effect. After all, if the status quo is in "Nash equilibrium," then it follows, by definition, that no individual has sufficient motivation to unilaterally deviate from it. So social change requires us to coordinate and move to a new equilibrium together. And ideology critique cannot help us to overcome such collective action problems.

The explanations offered by ideology critics and game theorists appear starkly at odds. The former explain the stability of bad social practices by reference to epistemic distortion, the latter by reference to problems of strategic interaction. But I will argue that there is no real conflict here. In fact, the very game-theoretic tools that show us why ideology critique is causally impotent in some cases also reveal causal mechanisms by which it may make a

---

term referring to any system of political ideas (Geuss 1981: 4-22). But I do not assume further that ideologies necessarily stabilize unjust regimes, serve the interests of an oppressor class, or anything of the like. Ideologies may very well do this, but it does not follow by my definition. Compare Sankaran (2020, 1443-1445).

significant difference in others. In particular, ideology critics point to two sorts of distortion that ideologies may engender. They may lead individuals to *misdescribe* the features of social outcomes, or to *reify* these features by seeing them as necessary rather than as contingent and changeable (e.g., Shelby 2003, 176-177; Jaeggi 2008, 65). An ideology critique that overturns these distortions can respectively change people's *preferences* or their *perceptions of the feasible set*. And it follows as a matter of basic game theory that ideology critique can therefore, in some cases, generate a new equilibrium that renders a previously insoluble collective action problem now soluble, and, in others, render our current arrangements no longer in equilibrium—thereby occasioning a shift away from the status quo all by itself. Recognizing this much also paves the way for more sophisticated game-theoretic (and evolutionary game-theoretic) analyses that explicitly incorporate ideology into their models of social stability and change.

2. Ideology Critique and its Critics

Unjust or otherwise suboptimal arrangements can persist for many reasons. For example, sometimes powerful individuals have a vested interest in maintaining the status quo, and others are powerless to change this. This situation is unfortunately common, but it raises no deep explanatory puzzle. More puzzling are cases where bad arrangements persist even though the individuals subject to them both seem to prefer some alternative outcome and to have the power to bring it about. For example, an oppressed majority may, in virtue of its sheer numbers, seem able to overthrow an oppressive minority. Why, then, doesn't the majority rise up? In such cases, ideology critics explain the persistence of bad arrangements by appeal to ideology: individuals must have a distorted understanding of the social world, or else they would change the status quo. Critiquing this ideology, and so clearing up this distortion, therefore appears a crucial task in promoting a better or more just world.

The project of ideology critique, however, faces at least three major worries. The first is epistemic (e.g., Celikates 2016, 4; Haslanger 2012, 22-30). Ideology critics aim to overturn distorted understandings of the social world shared by many individuals. In so doing, they claim a superior understanding to others, even—and paradigmatically—on issues where we typically defer to others, such as where their interest truly lie. This suggests that the perspective of the ideology critic is not only epistemically but morally fraught, since claiming that others do not know their own interests, or more generally that they only support the status quo because they are subject to epistemic distortion (rather than, say, because they genuinely and reasonably disagree with you), risks insult or disrespect (Heath 2000, 363). In responding to this worry, ideology critics face a heavy burden of proof.

A second worry concerns the ability of ideology critique to change hearts and minds. Grant that some population is subject to a distortive ideology. Widely shared beliefs are notoriously difficult to change. And ideologies are often thought to be especially sticky since they may involve not only beliefs but associated patterns of attention, tendencies to overrate or discount certain forms of evidence, and so on (e.g., Haslanger 2017). Ideologies, in other words, can involve not only distorted beliefs, but "belief immune systems" protecting those beliefs from change in the face of counter-evidence (Buchanan 2020, ch. 8). How, then, is a critic supposed to pierce through an ideology, and so cause the scales to fall from individuals' eyes? Here, too, the ideology critic faces a heavy burden in defending their approach as psychologically realistic.

These are both important worries, but they are not my focus here. Let us grant that ideology critiques are sometimes "successful" in the sense that they both successfully identify ideological distortion and successfully clear up this distortion in the minds of those previously subject to it. Still, even successful ideology critics face a third worry. In particular, the very

cases where ideology critique seems most needed—namely, those where some group seems to have the power to improve the status quo, but fails to do so—tend to share a common feature. This is that no individual can change the status quo on their own, so change requires many individuals to begin to act differently. But the need for such collective action raises notorious problems, which, some worry, undermine the causal efficacy of ideology critique: perhaps even a *successful* ideology critique cannot resolve collective action problems, and so cannot effect social change. As I have noted, this criticism is often framed in terms of game theory, and I will couch my discussion in the same terms—doing my best to explain everything in such a way that no familiarity with game theory (or, for that matter, ideology critique) will be required to follow along. I begin with the game-theoretic case against ideology critique, but in the end will argue that game theory can help us to understand not only the limits of ideology critique, but also its promise.

3. The Causal Inefficacy of Ideology Critique

A convenient entry point into the case against ideology critique is the game-theoretic staple of the Prisoner's Dilemma (Heath 2000, 366). The police have caught two partners in crime and have enough evidence to convict each of a minor offence, but can convict either of a serious offence only if their partner rats them out. If both stay mum, both get a short jail sentence; if both rat, both get a medium sentence. So, the police offer each a deal: if you rat your partner out and they stay mum then you get no jail time at all, but if your partner rats you out and you stay mum then you get a longer sentence than if you both rat.

If we assume, for now, that the prisoners are motivated only to minimize their jail time, then we can represent this game in the following matrix:

	Mum	Rat
Mum	2, 2	4, 1
Rat	1, 4	3, 3

Figure 1: Prisoner's Dilemma

Here, "4, 1" represents that the row player gets their 4th most preferred option while the column player gets their 1st most preferred option—where someone prefers A to B, in the relevant sense, if they are sufficiently motivated to bring about A rather than B given a choice between them. So each thinks: if my partner keeps mum, ratting gets me my most preferred outcome (no sentence) rather than my second most preferred outcome (a short sentence), and if my partner rats, ratting gets me my third most preferred outcome (a medium sentence) rather than my least preferred outcome (a long sentence), so I do better to rat no matter what. And thus, our prisoners both rat—even though each prefers that both stay mum.

The Prisoner's Dilemma is famous for its demonstration of how even perfectly rational, fully informed individuals may arrive at outcomes that all prefer to avoid. In fact, it is an especially troublesome case, because the collective action problem the prisoners face is *insoluble*. The problem is not that they are trapped at a suboptimal Nash equilibrium (in the sense that each would prefer some other equilibrium) and that they must coordinate to achieve a better one, but that they are trapped at the only equilibrium, since it is impossible for them to achieve any other outcome. Either will rat no matter what the other does.

To see the contrast here, let us transform the Prisoner's Dilemma into an Assurance Game (sometimes known as a "Stag Hunt") by modifying it slightly: each prisoner is now motivated not only to reduce their own jail time, but also somewhat to reduce the other's jail time, such that each now prefers the outcome in which each stays mum (and receives a small

sentence), to the one in which they rat (and receive no sentence) while the other stays mum (and receives a long sentence):[2]

	Mum	Rat
Mum	1, 1	4, 2
Rat	2, 4	3, 3

Figure 2: Assurance Game

Now, each most prefers to stay mum if the other stays mum, but to rat if the other rats. If either can be *assured* that the other will stay mum, then they will stay mum too, since each most prefers the outcome in which both stay mum. But if either predicts the other will rat, then they will rat too, since each still prefers to rat if the other rats. So both the outcomes in which each rat and in which each stay mum are in equilibrium, and the outcome our prisoners reach depends on whether each has sufficient assurance the other will stay mum.

In the Assurance Game, we again have a collective action problem, but one that is in principle *soluble*: if the prisoners can coordinate, say, through communicating, then they may succeed in achieving their preferred equilibrium. But if the suspects cannot communicate or more generally coordinate, then they will again end up at the suboptimal equilibrium. So we have two types of cases where suboptimal outcomes may arise due to collective action problems. In cases resembling the Prisoner's Dilemma, a suboptimal outcome is the only (or best) outcome in equilibrium, so it is not even in principle possible for individuals to solve

[2] It is a common misconception that game theorists assume individuals to be purely self-interested. Game-theoretic analyses do assume that people act to achieve their preferred outcomes, but this is compatible with them having any preferences over outcomes at all—including altruistic preferences or those that reflect moral commitments.

their collective action problem and achieve a preferred outcome through coordination. In cases resembling the Assurance Game, collective action is in principle possible because better equilibria are available. But no individual will unilaterally switch over, so in the absence of effective coordination the suboptimal equilibrium persists.

The connection to ideology critique is now coming into focus. As we have seen, ideology critics appeal to ideology to explain why bad arrangements persist even when individuals subject to them would seem to prefer some alternative outcome that they have the power to bring about. But these game-theoretic examples illustrate how suboptimal outcomes can arise and remain in Nash equilibrium—again, in the sense that no individual is sufficiently motivated to unilaterally deviate—in the absence of any epistemic defect. So, in many cases where such arrangements persist, the better explanation may be that individuals face a collective action problem. In such cases, ideology critique can at best succeed in convincing individuals that their current equilibrium is worse than some other outcome, but this will not change anything—any more than alerting the prisoners to their dilemma will lead them to stay mum. After all, as Heath puts it, "the mere recognition that the outcome is suboptimal does not change the incentives that each individual has to act in a way that contributes to it" (2000, 366). Instead, collective action is needed to bring about a change to the preferred outcome, and ideology critique cannot help individuals to solve this collective action problem.

For example, while it is tempting to think that oppressed majorities fail to overthrow oppressive minorities due to ideology, Rosen (1996) argues that a better explanation is often that the oppressed have structurally analogous preferences to an (*n*-person) Prisoner's Dilemma: each prefers the outcome in which everyone revolts to that in which no one revolts, but—given the high cost of revolting and the little impact any individual revolutionary

makes—each also prefers to "free ride" and not revolt regardless of what others do, so no one revolts (compare Tullock 1971):

	Revolt	Remain
Revolt	2, 2	4, 1
Remain	1, 4	3, 3

Figure 3: Revolution Game (Prisoner's Dilemma Preferences)

Or perhaps the oppressed instead have structurally analogous preferences to an (*n*-person) Assurance Game: each prefers to revolt if others revolt, but not to revolt if others don't:

	Revolt	Remain
Revolt	1, 1	4, 2
Remain	2, 4	3, 3

Figure 4: Revolution Game (Assurance Game Preferences)

In this latter case, both everyone revolting and no one revolting is in equilibrium. But if individuals lack the means to coordinate and so to provide one another with assurance that they will revolt—as is often the case under oppressive regimes—then they will end up at the suboptimal equilibrium where no one revolts. So, Rosen concludes, we can explain why oppressed majorities fail to revolt simply by pointing to the collective action problems they face. There is no need to posit a distorting ideology (1996, 505-509).

Sankaran (2020) goes further, arguing that even in cases where we grant that ideological distortion prevails, collective action problems still prevent ideology critique from dislodging bad conventions, since conventions are, on Sankaran's definition, in Nash equilibrium:

> It's vital to note that this collective action problem is a part of the formal properties of conventions, regardless of their content. This is the key feature of conventions that… ideology critics miss. [For example,] [e]ven if every member of a given

> community comes to see that the purity and chastity norms that organize practices surrounding women's sexuality are distorting and unjust, there is still no guarantee that agents will shift from the existing convention to a better one (Sankaran 2020, 11).

More specifically, Sankaran argues that "knocking a society from one convention to another requires more than just convincing everyone of the superiority of the new convention" (2020, 1439). After all, it follows from the formal definition of a Nash equilibrium that individuals will not unilaterally shift away from one, even if—thanks to a successful ideology critique—all come to see a new outcome as preferable to the status quo. So if ideology critique can only help people to recognize the suboptimality of the status quo, then it can do little to effect meaningful social change. Collective action problems appear to undermine the causal efficacy of ideology critique.

4. The Causal Efficacy of Ideology Critique

Rosen provides an illuminating analysis of why, in certain cases, oppressed majorities may fail to revolt against oppressive minorities. And Heath is surely right that, in many others, collective action problems rather than ideology similarly explain why bad arrangements endure. But there is a gap between this conclusion and Heath's claim that this is the better explanation of why "the vast majority of oppressive practices" persist (2000, 364). And while Sankaran's argument may appear to close this gap by showing that ideology critique cannot effect social change even when ideological distortion is present, this argument relies on a faulty premise: namely, that ideology critique can only show people that their current outcome is suboptimal. On the contrary, dispelling epistemic distortion can also lead individuals to see the motivationally-relevant features of the outcomes they are viewing differently, thus modifying their preferences. And it can lead them to see new options as available, thus

changing their perceptions of the feasible set. Since it is a matter of basic game theory that whether an outcome is in equilibrium depends on both variables, either sort of change can result in the creation of new equilibria or in the destruction of old ones. For example, ideology critique can transform insoluble Prisoner's Dilemma-style collective action problems (in which there are no preferable equilibria) into soluble Assurance Game-style collective action problems (in which there are), and it can even occasion social change without the need for any collective action at all.

To begin to explore these possibilities, recall that we can transform a Prisoner's Dilemma into an Assurance Game by swapping the ranking of each individual's top two options. In a Prisoner's Dilemma, each individual's most preferred outcome is ratting while the other stays mum and second most preferred outcome is both staying mum; in the Assurance Game these preferences are reversed. In particular, in our earlier rendition of the story, the difference was that in the former our suspects are motivated only to reduce their own sentences whereas in the latter each is also motivated to reduce their partner's sentence. Now, let us complicate the story: each prisoner is motivated to reduce the sentence of their friends but not their enemies, such that if they see each other as enemies, they will have Prisoner's Dilemma preferences, but if they see each other as friends, they will have Assurance Game preferences. Suppose, now, that to get the prisoners to confess, the police have fed each prisoner lies about the other to convince these erstwhile friends that they are in fact enemies. The prisoners are therefore subject to epistemic distortion, and this leads them to have Prisoner's Dilemma preferences such that the only equilibrium is both ratting. But if this epistemic distortion were cleared up, and each were to realize that the police had lied, then each would go back to seeing the other as their friend, and so would develop Assurance Game preferences. In that case, the outcome in which both stayed mum would become in

equilibrium, and the prisoners could perhaps coordinate to achieve it.

Clearing up epistemic distortion can therefore create a new equilibrium by changing individuals' preferences, and this can serve to destabilize equilibria through rendering coordination on a new equilibrium possible. But dispelling distortion can also destroy old equilibria, thus having a more radical destabilizing effect. To see this, let us begin, again, with the original Prisoner's Dilemma, but modify the distortion the police impose: this time, the police convince the prisoners that they will keep them safe, but each prisoner will actually be murdered by their mob boss if they rat. If this epistemic distortion were lifted, the prisoners (who, let us assume, are motivated above all to remain alive) would come to prefer either outcome in which they keep mum to either in which they rat:

	Mum	Rat
Mum	1, 1	2, 3
Rat	3, 2	4, 4

Figure 5: Loyalty Game

And the outcome in which each rats would therefore no longer be in in equilibrium. Instead, the only equilibrium would be each staying mum.

So far, we have considered cases where individuals have distorted beliefs about features of outcomes that produce preferences they would revise in the face of better information. And we have seen that clearing up such distortion can create new equilibria or destroy old ones, thus rendering collective action problems soluble or dissolving them altogether. But dispelling distortion can do more than just change individuals' preferences. It can also change their perception of the feasible set, which can, again, create or destroy equilibria. To see this, return once more to the original Prisoner's Dilemma in which individuals aim only to minimize their jail time (and care not at all about their counterparts).

One feature of this case is that the suspects each perceive themselves as having only two options: ratting or staying mum. But suppose now that the police have again distorted their beliefs, and hid that they actually have a third option: to call a lawyer who will get all charges dismissed on procedural grounds (in which case, we may further suppose, the other suspect will no longer receive any benefit from ratting). Failing to recognize this, the suspects face a Prisoner's Dilemma. But if they were to appreciate this third option, they would instead play:

	Mum	Rat	Lawyer
Mum	2, 2	4, 1	2, 1
Rat	1, 4	3, 3	2, 1
Lawyer	1, 2	1, 2	1, 1

Figure 6: Lawyer Game

Here, the outcome in which each rats is no longer in equilibrium. Instead, the only equilibrium is that each calls a lawyer. So modifying individuals' perceptions of the feasible set through clearing up epistemic distortion can also destroy and create equilibria—even without affecting their preferences over the options they already recognized.

Of course, these are not the sorts of cases that ideology critics have in mind. But I have taken the time to work through them here as a warmup for the main event, since, as I will now show, the very same causal mechanisms may obtain in such cases as well. Ideology critics, recall, hold that ideologies can distort our understanding of our social world by leading us either to *misdescribe* or to *reify* its features (e.g., Shelby 2003, 176-177; Jaeggi 2008, 65). In the former case, an ideology critique may lead individuals to a more adequate understanding of the motivationally-relevant features of outcomes, thereby changing their preferences. In the latter, it may lead them to recognize that other options are available, thereby changing their perception of the feasible set. And, just as in the examples we have considered, either sort of

change can create or destroy equilibria, thus partially or fully destabilizing the status quo.

With respect to the mere creation of equilibria, consider again Rosen's discussion of the oppressed majority. Suppose, to elaborate upon this case somewhat, that the oppressive minority has instilled an ideology suggesting that they are far more powerful than they actually are, and that this explains why the members of the oppressed class face an (*n*-person) Prisoner's Dilemma in which each prefers to stay home no matter what others do, even though each also prefers the outcome in which all revolt to that in which no one does (as in Figure 3). One can readily understand why everyone would most prefer to "free ride" when the alternative is going up against a mighty power. If this distortion is lifted thanks to a successful ideology critique, individuals may realize that the cost of collective revolting is not so high, and so come to face an (*n*-person) Assurance Game instead, in which their most preferred outcome now involves them revolting on the condition that others do so as well (as in Figure 4). This puts revolting in equilibrium for the oppressed individuals, and so renders their previously insoluble collective action problem now soluble (on the condition that they are able to communicate or otherwise assure one another that they will join in).

Of course, there is no guarantee that ideology critique can have this effect in all cases. Sometimes, even after all distortion is lifted, people may still face an insoluble collective action problem. For example, suppose that whether each individual prefers revolting or staying home on the condition that others revolt depends on two factors: the extent to which they view revolting as affecting their own interests, and the extent of their other-regarding concern as determined by their altruism, reciprocity, or solidarity. While the above ideology critique may lead an individual to see the personal cost of revolting as lower than they previously thought, this will only succeed in transforming the collective action problem from soluble to insoluble if the cost of revolting (on the condition that others revolt) is driven low enough that it is

outweighed by the individual's other-regarding concern. Likewise, it may be possible for another sort of ideology critique to transform this collective action problem by increasing other-regarding concern, for example, if, as above, individuals categorize others into different groups (say, friends and enemies), have differential concern for those in different categories, and they are subject to an ideology that leads to miscategorization. But once again, there is no guarantee that this latter sort of ideology critique can increase other-regarding concern enough to succeed in reversing individuals' preferences either. So the point here is not that game theory shows us that ideology critique *always* works. It is rather that analyzing these situations along game-theoretic lines helps us to understand the conditions under which ideology critique is, and is not, causally efficacious.

Another way ideology critique can cause social change by altering people's preferences or perceptions of the feasible set has occurred in many real-world cases involving the abolition of the sort of practices on which Sankaran focuses, such as female genital mutilation and footbinding. While there is still the need to coordinate action away from these practices—since unilateral deviation typically has very high costs when one is dealing with practices such as these, where parents or their children are subject to severe enforcement if they deviate from the norm—their stability often relies on false empirical beliefs about their effects (whose correction tends to result in a change to individuals' preferences), and on assumptions that such practices are "natural" and unalterable (whose correction leads individuals to recognize a more expansive feasible set). Perhaps unsurprisingly, then, dispelling both sorts of ideology has historically played an important role in generating new equilibria, paving the way for collective action away from these practices—for example, through the use of carefully planned

and executed pledges to stop engaging in and enforcing these practices.[3] Notably, while such cases involve the creation of a new equilibrium, the pathway by which ideology critique effects change in them is different than in the previous case. It doesn't transform an insoluble collective action problem into a soluble one, but rather generates a soluble collective action problem where, previously, no collective action problem existed—since, prior to the ideology critique, people (given their false beliefs) preferred the status quo over all alternatives they perceived as feasible.

Let us now turn to two highly stylized cases in which ideology critique may not merely create new equilibria but also destroy old ones, thus generating a shift away from the status quo all by itself. Suppose first that a society formally permits women to do whatever work they please, but an ideology prevails in which women can only achieve happiness through homemaking and other forms of domestic labor. When deciding who will engage in wage labor in the formal economy and who will engage in domestic labor at home, a heterosexual couple might therefore find themselves playing the following game:

		Wife: Wage	Wife: Domestic
Husband	Wage	2, 2	1, 1
	Domestic	3, 3	4, 4

Figure 7: "Happy Housewife" Game

Here, the only equilibrium is that husband engages in wage labor and the wife engages in domestic labor. Troublingly, this equilibrium may persist even if the wife is deeply unhappy,

[3] Sankaran (2020, 1450-1452) himself draws on Mackie (1996; 2000) here, who provides a helpful overview of the empirical literature.

because she buys into the ideology that she will be even more miserable if she enters the formal workforce, so that she prefers (or, equivalently, is sufficiently motivated) to avoid this outcome. Perhaps even more troublingly, if this same game is played throughout society in a large number of households, it may lead, among other things, to widespread patterns of economic dependence—even if this is an outcome that no one aims to promote.

Suppose now that the ideology is lifted: the wife reads Betty Friedan's (1963) "Myth of the Happy Housewife" and comes to recognize that she would in fact be happier entering the formal workforce. Everything else remains the same, but given her motivation to achieve happiness, she comes to most prefer the option where both she and her husband engage in wage labor:

		Wife: Wage	Wife: Domestic
Husband	Wage	2, 1	1, 2
	Domestic	3, 3	4, 4

Figure 8: "Happy Housewife" Game after Ideology Critique

Now, the status quo of the husband working for a wage and the wife doing domestic labor is no longer in equilibrium. Instead, the outcome in which both take a wage is the only equilibrium. So the ideology critique succeeds, through changing the wife's preferences, in changing the status quo. And if this critique prevails throughout society, many other women may begin to make similar choices, leading to more fulfilling lives and less entrenched patterns of economic dependence. No collective action problem prevents this social change.

The point of this example is to illustrate how, in principle, social change need not require collective action. Sometimes, many individuals each have the opportunity to make a choice largely independently of how others' choose, but the cumulative effects of all such

choice is a change to widespread social patterns. Ideology critique can be especially potent in such cases. But, as I have noted, the above example is highly stylized, and real life is not so simple. For one thing, women differ considerably in how they like to spend their time. Some may prefer domestic labor to working in the formal economy regardless of any ideology they are subject to. For another, there are often many factors preventing women from entering the formal workforce. They may face significant social pressure to stay home from their husbands and from broader society when most other women still occupy this social role, and the workplace may be highly unwelcoming for women who remain a small minority. The above example ignores all this and so might seem to obscure the fact that, in any realistic case, women face an Assurance Game-style collective action problem amongst themselves: even if an ideology critique leads them to prefer to enter the formal workforce on the condition that enough other women do so as well, they may still prefer not to be first movers.[4] This suggests that an ideology critique may be unable to dissolve the equilibrium in the "Happy Housewife" Game in the way I have proposed, at least for most couples. Once we add a dash of realism, we find that collective action is needed to occasion widespread social change after all.

But this is too quick. Collective action is not the only thing that can solve a first mover problem. So, too, can heterogeneity in agents' preferences. In particular, suppose that due to differences in how women like to spend their time, the different social pressures they face, and so on, women differ in their threshold of how many other women must enter the formal workforce before they prefer to do so themselves. Some prefer to enter once 30% of other women do so, but others prefer not to enter it until 40% or 50% enter it, and so on. Under

[4] As we will see in the next section, it is also possible to model the gendered division of labor as arising from a *bargain* between a household's members over the distribution of labor.

such conditions, a few women deciding to enter the formal workforce has the potential to trigger a cascade whereby many others follow suit. This will occur if there is some "tipping point" of, say, 30% workforce participation, such that once this threshold is hit, several women who have a 30% threshold will also join, causing several women with a 31% threshold to also join, causing several women with a 32% threshold to join, and so on, until workforce participation is considerably increased.

This example remains highly simplified, but it nevertheless illustrates the sort of tipping point and cascade dynamic that is now widely recognized as a central mechanism of social change by contemporary social scientists (e.g., Kuran 1995, 71-73; Sunstein 1996; Bicchieri 2017, 163-194). An appreciation of this dynamic is important, because it suggests that even if an ideology critique only succeeds in transforming the "Happy Housewife" Game for a few trendsetters, this may be enough to occasion widespread social change—especially if the critique also alters the preferences of other women by decreasing the threshold percentage of women participating in the workforce it would take for them to prefer to join as well. In other words, if an ideology critique destroys the equilibrium of the husband taking a wage and the wife engaging in domestic labor for even a small minority of couples, this may be enough to trigger a cascade that destabilizes the broader social order. Of course, there is no guarantee that this will occur—tipping points don't always exist close enough to the status quo that an ideology critique can push society beyond them—but when it does, this can lead to significant social disruption. Crucially, no collective action is needed to effect this change.

As a final example, suppose that politicians are debating how to remedy high rates of poverty and crime in a predominantly Black neighborhood. The politicians accept an ideology of "racial essentialism" on which it is an essential feature of Black people that they are less productive and more prone to crime, or perhaps a more modern cultural analogue of this view

on which such racial differences arise due to deep aspects of Black people's culture that will remain stable in the face of any potential intervention (compare Shelby 2003, 177). So they believe that the only two relevant options are a welfare bill aimed at ameliorating poverty and a "tough on crime" bill aimed at reducing crime. The policy will only pass given bipartisan support, and one party leader must first propose a bill, which the opposing party leader must then accept or reject. The proposer prefers the welfare bill passing, to nothing passing, to the crime bill passing; the other has the reverse preferences. They therefore face the following sequential game:

	Accept Bill	Reject Bill
Propose Welfare Bill	1, 3	2, 2
Propose Crime Bill	3, 1	2, 2

Figure 9: Policy Game (row goes 1st; column goes 2nd)

Here, the only equilibrium is that the welfare bill is proposed but rejected: since the proposer goes first, they will not propose the crime bill knowing that the other party will accept it (the proposer's third choice), when they can instead propose a welfare bill that the other party will reject (the proposer's second choice). Gridlock prevails.

Suppose, now, that the ideology is lifted: our politicians realize that the conditions in the neighborhood are due not to essential features of its members or their culture, but rather to a contingent lack of access to meaningful social and economic opportunities. A third option thus enters the policy space: a bill involving a suite of policies aimed at expanding such opportunities. Each politician sees this bill as second best, but otherwise has the same preferences as before:

	Accept Bill	Reject Bill
Propose Welfare Bill	1, 4	3, 3
Propose Crime Bill	4, 1	3, 3
Propose Opportunity Bill	2, 2	3, 3

Figure 10: Policy Game After Ideology Critique (row goes 1st; column goes 2nd)

Now, the outcome in which the welfare bill is proposed and rejected is no longer in equilibrium: since the proposer goes first, they will not propose the welfare bill knowing the other party will reject it (the proposer's third choice) or the crime bill knowing that the other party will accept it (the proposer's fourth choice), when they can instead propose an opportunity bill the other party will accept (the proposer's second choice). Instead, the only outcome that is in equilibrium is the opportunity bill being proposed, accepted, and passed. So, once again, an ideology critique can succeed in effecting real social change—this time, through altering the feasible set .

It bears worth emphasizing one last time that—as this last example will surely highlight for many—there is no guarantee ideology critique will lead to meaningful change in any given real-world case, since there is no guarantee that the underlying strategic circumstance or "game" resembles one of our examples. As Heath (2001, 366-367) rightly points out, some good evidence that we are in a situation where an ideology critique is unlikely to occasion change, since we are instead trapped in a collective action problem (or, we might add, a situation where those who wish to preserve the status quo are powerful enough to do so) is that nearly everyone has already seemed to internalize the critique, "but nothing ever changes." Such evidence is not conclusive, however, especially given the possibility of tipping points: it could be that nothing *seems* to be changing, but people's underlying attitudes are indeed changing and bringing us closer and closer to a tipping point where rapid change will suddenly

occur (compare Bicchieri 2017, 192). In any event, the point, as always, is that game theory can help us to model and so understand the conditions in which ideology critique can effect social change, not that we are always in such conditions.

We have now examined a number of cases in which ideology critique can, through clearing up epistemic distortion, destabilize the status quo. These examples suggest that, contrary to what critics of ideology critique argue, ideology critique can sometimes be causally efficacious after all. But before closing this discussion, let us consider a few objections, as well as some possible extensions to this analysis that our responses to these objections will suggest.

5. Objections and Extensions

The first objection is a somewhat technical one. The central argument of this paper has relied on the idea that a successful ideology critique can change people's preferences or perceptions of the feasible set. But this might seem like a slippery way of talking, and one that is rather alien to standard game-theoretic analyses. Can we make it more rigorous?

One way to do this is to employ the notion of a "perspective." An individual's perspective includes the various schemata or filters that mediate their perception of the external world; formally, it is a mapping from the external world to an "internal language" or other representational system (Page 2007, 31; Muldoon 2016, ch 3). For example, from one person's perspective, the world might contain immaterial souls; from another, it contains only physical objects. Similarly, from one perspective, certain aspects of our social world (such as social hierarchy) may be natural and unchangeable, whereas from another these aspects can be modified or replaced. Rather than modeling individuals as having preferences over *actual* states of the external world, and specifically those states that are *actually* feasible, we can now model them as having preferences over states of the world *as encoded by their perspective*, and *that*

are deemed feasible by their perspective.

This perspectival framework for modeling preference and choice can be formalized in a highly rigorous way and is, in fact, independently motivated given its ability to solve various puzzles in the formal theory of rational choice (Kogelmann 2018). And it provides a straightforward analysis of what is going on in cases of ideology critique: we can understand ideologies as collectively held perspectives that misrepresent the world, and ideology critique as helping to clear up such misrepresentations, leading individuals to adopt more accurate perspectives. So, in this framework, ideology critique does not strictly speaking change people's preferences over unchanging objects, but rather changes their perspectives and so the objects over which they have preferences (and which objects they perceive as feasible). To return to an earlier example: a woman might always prefer to lead a happier lifestyle, such that what changes after a successful ideology critique is not this preference, but rather her perspective on what her options are. Maybe her previous perspective excluded being happy in the formal workforce as a feasible option, but her new perspective includes it. Notably, this perspectival framework not only allows us to formalize our analysis to this point, but also suggests further modeling approaches that may help us better to understand the prospects for ideology critique. For example, rather than modeling the ideology critic as exogenous (as we have done so far), we might construct models or simulations in which one action available to agents is to engage in an "ideology critique" that (perhaps with some probability) alters others' perspectives, so that we can observe what dynamics unfold.

Second, and perhaps more troublingly, one might protest that, formal rigor aside, I have invoked tremendously simplified examples: in the real world, social change is much more complicated than I have made it seem. This is certainly true. But the simplified cases I have examined are nevertheless valuable because they have allowed me to isolate and so illustrate a

number of causal mechanisms—in particular, how changing preferences or changing perceptions of the feasible set can result either in the creation of new equilibria or the destruction of old ones—by which ideology critique can destabilize the status quo in the real world. And I have furthermore chosen them to illustrate actual contexts in which such mechanisms plausibly activate. For example, in some contexts, social arrangements—say, those involving relations of economic dependence—may persist due to widespread patterns of individual choice, which ideology critique may suffice to change, especially given the possibility of tipping points and cascades. In others, we already have institutions in place to solve collective action problems—say, democratic politics—such that ideology critique can effect social change without running into them. My hope is that my admittedly simplified examples will pave the way for far more sophisticated analyses that make use of both the tools of game theory and the insights of ideology critics, whose models of social stability and change, I have tried to show, are far more complementary than either camp appears to have realized.

That said, I do not mean to suggest that the only causal mechanisms by which ideology critique may destabilize the status quo must involve changes to preferences or perceptions of the feasible set; these are just the most salient mechanisms given ideology critics' focus on ideologies that lead us either to misdescribe or to reify features of the social world. For example, ideology critique may also destabilize the status quo through remedying people's false perceptions of *others*' preferences. The most obvious case of this sort will arise when individuals who would otherwise face a soluble Assurance Game-style collective action problem are under the false impression that others have insoluble Prisoner's Dilemma-style preferences, rendering assurance impossible (since those with Prisoner's Dilemmas preferences will never engage in collective action). Indeed, to really render a collective action problem soluble, an ideology critique must therefore not only change preferences or

perceptions of the feasible set, but also render this change common knowledge—as I have implicitly assumed occurs in cases of a successful ideology critique throughout. So, to head off another objection, nothing in my analysis requires us to assume common knowledge of preferences. I have just assumed this for simplicity, again, to isolate relevant causal mechanisms.

Our next objection zooms in on a particular modeling choice I have made throughout this paper. I have focused on "one-shot" games in which individuals interact a single time, instead of employing a more dynamic analysis involving repeated games or evolutionary game theory. But wouldn't a dynamic analysis be more appropriate, given that ideologies tend to shape actions and social arrangements over long periods of time, rather than all at once?

As usual, my choice here has been driven primarily by my aim of most simply and cleanly isolating mechanisms by which ideology critique can effect social change. And since I have primarily focused on cases where an ideology critique might cause some group, once and for all, to engage in collective action to change the status quo, this one-shot analysis is not only simpler but perfectly appropriate. Whereas modeling the long-term effects of ideology on the development of norms, conventions, or equilibria more generally would certainly require a more dynamic analysis, my basic approach has instead been to take as given that we have already reached a point where certain social arrangements are stable, and to ask how an ideology *critique* might destabilize this status quo. Nevertheless, more dynamic analyses may very well reveal further mechanisms by which ideology critique can effect social change, and here evolutionary game theory may be especially illuminating.

Evolutionary game theory differs from classical game theory in how we model agents' decisions and how we interpret the "payoffs" or numbers in our matrices. In classical game theory, we attribute agents preferences over outcomes and assume that they will always adopt

a "best response" strategy of choosing the outcome they most prefer (or that gives them the highest payoff), conditional on the strategies others adopt. In evolutionary game theory, we permit agents to employ a much wider range of strategies, and the payoffs that agents achieve playing those strategies correspond to how likely those strategies are to spread through the population in future rounds—for example, because strategies with higher payoffs correspond to strategies that others are more likely to regard as "successful" and so to copy. This requires us to specify, among other details, not only the underlying game agents are playing, but also some rule by which agents are matched with others each round (for example, they may be paired randomly), as well as the transmission rule by which strategies are made more or less likely to spread among the population across time.

One major advantage of evolutionary game theory is that it lets us investigate not only the equilibrium outcomes of games, but also which equilibria are more or less likely to arise over time through the repeated interactions of agents. Here, a key idea is a "basin of attraction," where equilibria with larger basins are those that populations will evolve to from a wider range of initial strategy distributions. Status quo equilibria with larger basins of attraction are also harder to disrupt, because it takes a larger number of agents acting against the status quo to dislodge them (O'Connor 2019, ch. 9).

This last point is crucial for our purposes. Since the size of a basin of attraction depends both on people's preferences and their perceptions of the feasible set, and we have seen that ideology critique can change both, evolutionary game theory may therefore allow us to model another, subtler mechanism, by which ideology critique can destabilize the status quo. Namely, it can render collective action possible, or easier to achieve, by reducing the number of individuals needed to engage in collective action in order to effect change. And this suggests an especially fruitful line of further research, given that, in recent years, evolutionary

game theorists have made much progress exploring how unequal or discriminatory conventions evolve (e.g., Bruner 2017; O'Connor 2019; O'Connor, Bright, and Bruner 2019). For example, they have constructed much more sophisticated models of the household division of labor than that suggested above, in which couples' decisions are modeled as a bargain over how to determine the distribution of labor, playing out repeatedly across many households and many generations (O'Connor 2019). But as far as I am aware, they have not yet examined how the evolution of such conventions might be redirected in the face of an ideology critique, or how such a critique might reduce the basin of attraction surrounding an unequal convention.

Two final objections challenge the significance of my analysis rather than its rigor or appropriateness. First, one might object that the core argument of this paper is trivial. By definition, a Nash equilibrium is an outcome from which no individual prefers to unilaterally deviate. So *of course* changing preferences or perceptions of the feasible set can create or destroy equilibria: whether or not some outcome is in equilibrium is a function of precisely such variables. This is true, but I have nevertheless carefully worked through various examples that illustrate this point because while it is indeed trivial that changing individuals' preferences or their perceptions of the feasible set can change which outcomes are in equilibrium, it is far from trivial to point out that ideology critique can change preferences or feasible sets in this way, thereby altering the status quo. Indeed, this is precisely what critics of ideology critique have overlooked in their efforts to downplay the causal efficacy of ideology critique, since they have assumed that the point of ideology critique is to help people see what games they are already playing rather than to modify such games. This misunderstanding is understandable given that ideology critics do not generally put things in terms of preferences, feasible sets, Nash equilibria, and the like—it takes some effort to show, as I have tried to do here, how the

language of ideology critique can be translated into the language of game theory. But once we bring such connections into focus, we again find grounds for interdisciplinary reconciliation rather than conflict.

Finally, does this intertranslatability show that ideology critique is useless after all, since ideology critics merely "reinvent the wheel" (Sankaran 2020)? I do not see why it should. In the first place, the fact that two research programs are compatible does not show that only one is valuable, given divergence in their methodologies, points of emphasis, or central concerns. Although I have argued that everything ideology critics want to say is compatible with game theory, the point remains that game theorists rarely explicitly model ideology. The insights of ideology critics can enrich our game-theoretic modeling and our understanding of social change and stability more generally, in the ways I have suggested—as can the insight of game theorists facilitate the project of ideology critique. In the second place, game theorists and ideology critics are in one sense engaged in very different projects, and self-consciously so. The former are primarily engaged in an explanatory project, and the latter in a practical or "ameliorative" project. To adapt Marx (1998): game theory may help us to model and interpret the world, in various ways, but the point of ideology critique is to change it.

6. Conclusion

I have shown that the same game-theoretic tools that critics of ideology critique employ to argue against the causal efficacy of ideology critique can be repurposed to reveal causal pathways by which ideology critique can effect social change. Collective action problems are important, and we should not assume that ideology critique is always sufficient to make a meaningful difference in the face of them. But nor should we assume that ideology critique is always causally inefficacious either. It is a misapplication of game theory to think that the

strategic circumstances or "games" we encounter are given—that our preferences and perceptions of the feasible set are fixed rather than a function of our beliefs about the world, which are all too often subject to ideological distortion. And it is a misunderstanding of ideology critique to think that it can only help us to see the structure of the games we are already playing. Instead, the point of ideology critique is to change the games we face by altering our preferences or perceptions of the feasible set. This can create or destroy equilibria (or, perhaps, shrink their basins of attraction) thus partially or fully destabilizing the status quo.

Acknowledgments

For helpful comments and discussion, I would like to thank Brian Berkey, Justin Bruner, Allen Buchanan, Jerry Gaus, Laura Martin, Alexander Motchoulski, Sarah Raskoff, Kirun Sankaran, and two anonymous referees. Thanks also to audiences at the PPE Society Fifth Annual Meeting in New Orleans and the 2022 APA Central Division Meeting in Chicago.

Works Cited

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

Bruner, Justin P. 2017. "Minority (dis)advantage in population games." *Synthese* 196: 413-427.

Buchanan, Allen. 2020. *Our Moral Fate: Evolution and the Escape from Tribalism*. Oxford: Oxford University Press.

Celikates, Robin. 2006. "From critical social theory to a social theory of critique: on the critique of ideology after the pragmatic turn." *Constellations* 13: 21–40.

Friedan, Betty. 1963. *The Feminine Mystique*. New York: W. W. Norton & Co.

Geuss, Raymond. 1981. *The Idea of a Critical Theory*. Cambridge: Cambridge University Press.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Haslanger, Sally. 2017. *Critical Theory and Practice.* Amsterdam: Koninklijke van Gorcum.

Heath, Joseph. 2000. "Ideology, irrationality and collectively self-defeating behaviour." *Constellations* 7: 363–371.

Heath, J. 2001. "Problems in the theory of ideology." In J. Bohman & W. Rehg (Eds.), *Pluralism and the Pragmatic Turn: The Transformation of Critical* Theory, 163-190. Cambridge: MIT Press.

Jaeggi, Rahel. 2008. "Rethinking ideology." In B. de Bruin & C. F. Zurn (Eds.), *New Waves in Political Philosophy*, 63–86. New York: Palgrave McMillan.

Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge: Harvard University Press.

Mackie, Gerry. 1996. "Ending footbinding and infibulation: A convention account." *American Sociological Review* 61: 999–1017.

Mackie, Gerry. 2000. "Female genital cutting: the beginning of the end." In B. Sheil-Duncan & Y. Hernlund (Eds.), *Female "Circumcision" in Africa: Culture, Controversy, and Change*, 245–282. Boulder: Lynne Reinner Publishers.

Marx, Karl. 1998. *The German Ideology: Including Theses on Feurerbach and Introduction to the Critique of Political Economy.* Amherst: Prometheus Books.

Kogelmann, Brian. 2018. "What we choose, what we prefer." *Synthese* 195: 3221-3240.

Muldoon, Ryan. 2016. *Social Contract Theory for a Diverse World: Beyond Tolerance.* New York: Routledge.

O'Connor, Cailin. 2019. *The Origins of Unfairness*: *Social Categories and Cultural Evolution*. Oxford: Oxford University Press.

O'Connor, Cailin, Liam Kofi Bright, and Justin P. Bruner. 2019. "The Emergence of Intersectional Disadvantage." *Social Epistemology* 33: 23-41.

Page, Scott E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton: Princeton University Press.

Rosen, Michael. 1996. *On Voluntary Servitude: False Consciousness and the Theory of Ideology*. Cambridge: Polity Press.

Stanley, Jason. 2015. *How Propaganda Works*. Princeton: Princeton University Press.

Sankaran Kirun. 2020. "What's new in the new ideology critique?" *Philosophical Studies* 177: 1441-1462.

Shelby, Tommie. 2003. "Ideology, racism, and critical social theory." *The Philosophical Forum* 34: 153–188.

Shelby, Tommie. 2016. *Dark Ghettos: Injustice, Dissent, and Reform*. Cambridge: Harvard University Press.

Sunstein, Cass R. 1996. "Social Norms and Social Roles." *Columbia Law Review* 96: 903-968.

Srinivasan, Amia. 2016. "Philosophy and Ideology." *Theoria* 31: 371-380.

Tullock, Gordon. 1971. "The Paradox of Revolution." *Public Choice* 11: 89-99.