Partisanship and The Trolley Problem: Understanding Republicans’ and Democrats’ willingness to sacrifice members of the other party

Michael Barber
Brigham Young University
barber@byu.edu
Ryan Davis
Brigham Young University
rdavis@byu.edu

Abstract

Do partisans view members of other party with lower moral status? While research shows they view out-partisans quite poorly, we show that affective polarization extends to expressing a willingness to sacrifice an out-partisan’s life. We report the first study to consider partisanship in the classic ‘trolley problem’ in which respondents are asked whether they would sacrifice an individual’s life in order to save the life of five individuals. We explore this issue with a nationally representative survey experiment in the United States, inquiring about politicized variants of the trolley problem case. First, we vary the political affiliations of both the group of five (to be saved by turning the trolley) and the single individual (to be sacrificed by turning the trolley). We find that individuals are less willing to sacrifice a co-partisan for the sake of group of out-partisans. These findings go beyond earlier work by suggesting that partisans not only hold negative attitudes and judgements toward political out-groups, but that they will at least signal approval of differing moral treatment. We take stock of how these results bear on normative questions in democratic theory.
What moral obligations do Americans recognize toward their political opponents? Regarding fellow citizens – including political adversaries – as worthy partners for cooperation has long been thought to be an important basis for democratic governance. Contemporary philosophers, with a nod to Aristotle, label this value as civic friendship Rawls (2005); Cohen (1989); Ebels-Duggan (2010); Vallier (2019).

Affective political polarization poses an intuitive challenge to realizing the values associated with political cooperation among adversaries. If people dislike members of the other party and want to avoid contact with them Huber and Malhotra (2017); Iyengar and Westwood (2015); McConnell et al. (2018), then their political attitudes dim the prospects for realizing civic friendship. Not only have partisans become more polarized Bishop (2009); Hetherington and Rudolph (2015); Mason (2018), but this trend has primarily manifested through increasing hostility toward members of the other party Abramowitz and Webster (2016).

Antagonistic feelings may also shape how individuals think about the moral status of members of opposing parties. Extreme affective polarization is associated with a greater willingness to dehumanize political opponents, ascribing to them properties of being less evolved or more animal like Martherus et al. (2019); Cassese (2019). A willingness to dehumanize predicts greater authoritarianism, partisan bias, and reduced tolerance Martherus et al. (2019); Cassese (2019). Indeed, the COVID-19 pandemic has sadly illustrated this with partisan commentators suggesting deaths in ‘Blue states’ are less concerning than those in ‘Red states’.1 Furthermore, recent unrest in US cities has shown political opponents engaging in lethal violence.2

Prior work has focused on whether partisans will predicate negative traits of political opponents. We are interested in whether negative affect extends to approving of different kinds of moral treatment. We explore this issue with a nationally representative survey experiment, inquiring about politicized variants of the ‘trolley problem’ case. First, we vary the political affiliations of both the group of five (to be saved by turning the trolley) and the single individual (to be sacrificed by turning the trolley). As in other survey applications of the trolley case, professed willingness to save the greater number is high. However, we find that individuals are less willing to sacrifice a co-partisan for the sake of group of out-partisans. Second, we consider how much this reflects negative partisanship by comparing partisanship with a variety of other out-groups. Among those considered, only neo-Nazis received less support by respondents than opposing partisans, and then only among Democrats. For Republicans, all out-groups were indistinguishable. These findings go beyond earlier work by suggesting that partisans not only hold negative attitudes and judgements toward political out-groups, but that they will at least signal approval of differing moral treatment. In the final section, we take stock of how these

---

1https://twitter.com/Acyn/status/1243004099206893583?s=20, accessed April 6, 2020
results bear on normative questions in democratic theory.

In the simplest version of the trolley problem, one must decide whether to pull a lever, which will bring about the death of one person in order to avoid the deaths of the five. Beginning with its introduction by Foot (1967), the trolley problem has prompted a great deal of debate in philosophy and psychology Thomson (1976, 2008); Christensen and Gomila (2012).

Most respondents act to bring about the utility promoting outcome Navarrete et al. (2012). There are two significant classes of exceptions. First, the more an agent must monitor their intentional activity so as to secure the death of one person, the more objectionable such acts feel. For example, subjects are more willing to accept pulling a lever to save five at the expense of one than they are to accept pushing one person in front of the oncoming trolley in order to save five Lanteri, Chelini and Rizzello (2008); Greene and Haidt (2002); Waldmann and Dieterich (2007). Second, individuals are less willing to sacrifice a person with whom they share an identity, such as a relative or romantic partner Bleske-Rechek et al. (2010).

The rise of negative partisanship is well documented in the political science literature. Relative to the past, contemporary partisans use fewer positive traits to describe opposing candidates and experience greater negative emotion when thinking about them Iyengar and Krupenkin (2018). Although warm feelings toward co-partisans have not correspondingly increased, antagonism toward political adversaries has been associated with greater party loyalty and heightened political participation Abramowitz and Webster (2016); Iyengar and Krupenkin (2018). Beyond attitudinal measures, negative partisanship predicts biases in behavior in the lab as well as in practice. In economic games in the lab, participants favor co-partisans and impose penalties on opposing partisans Iyengar and Westwood (2015); Carlin and Love (2013). Survey experiments show respondents avoid offering scholarships or jobs to opposing partisans, even when more qualified than in-party alternatives Iyengar and Westwood (2015); Gift and Gift (2015), and also more willing to date co-partisans Huber and Malhotra (2017).

There is reason to think that negative affective partisanship might extend to moral judgments. Cassese Cassese (2019) finds that members of both major political parties regard political out-group members as less human, both in a willingness to apply more ‘animalistic’ traits to out-group members, and also willing to associate them with ‘less evolved’ visual representations. Dehumanization at least points in a morally worrisome direction, in particular to potential consequences for compromise, gridlock, and a willingness to believe false or misleading media about opponents Leidner, Castano and Ginges (2013).

The literature on dehumanizing opposing partisans stops short of any evidence for the conclusion that
partisans will hold differing moral judgments about each other. However, there is evidence that dehumanization predicts violence and other norm violations in other contexts Cassese (2019); Viki, Osgood and Phillips (2013); Kouchaki et al. (2018). Kalmoe (2014) has recently found that some partisans are motivated by violent language, focusing on the use of metaphors comparing political activity to war, fighting, or other forms of conflict. Taken together, these results appear to support the expectation that partisanship will affect moral judgments. Yet we are unsure if partisans use dehumanizing language and violent metaphors because they think differently about the moral status of political opponents, or merely for rhetorical or motivational purposes.

The trolley problem provides a well-established way of assessing moral judgments about members of different groups Swann Jr et al. (2010). Although responses to the trolley problem do not directly predict behavior Bostyn, Sevenhant and Roets (2018), they do help answer questions about inter-group moral judgements that meaningfully differ both from expressions of affect and from the other behaviors associated with negative affect. Moreover, if negative affect and moral judgment do align, that would lend additional, novel support for predicting that behavioral norm violations observed with dehumanization in other contexts might also apply to the dehumanization of political enemies.

These differences with literature enable our study to build on earlier work with respect to three questions:

First, does the affect characteristic of negative partisanship extend to moral judgment? Are partisans expressing only enmity or ill feeling toward their opponents, or do they have, in addition, a moral intuition that out-party members can be treated differently? Because it is conceptually possible to hold strongly negative affect toward persons to who one at the same time imputes a common moral status, moral judgments are in an important respect more serious than demonstrations of affect. Even if survey reports don’t represent intentions, they do appear to represent genuine attitudes, in contrast with merely expressive attitudes. Berinsky (2018); Gerber and Huber (2009)

Second, is negative partisan affect reactive or objective? Philosophers distinguish between reactive and objective affective emotions Darwall (2006). Reactive emotions are directed at another person, while objective motions respond to a state of affairs. It makes sense, for example, to feel angry at or grateful to a person. In contrast it sounds infelicitous to be happy at or to a person. The difference is not merely linguistic. Reactive emotions play a distinctive functional role in an agent’s psychology Macnamara (2011). Some political scientists have wondered whether negative affect might be directed only at the circumstance in which politics is discussed, rather than toward opposing partisans in particular Klar, Krupnikov and Ryan (2018). If affective polarization is about disliking partisanship rather than reacting against partisans, we should not expect departures
from standard utilitarian responses to the trolley problem when partisanship is introduced.

Third, if negative partisanship does represent a judgment about moral status, what is the relevant comparison class for this judgment? In other words, if partisans do treat out-party members morally differently, how great is the difference? What are other kinds out-groups that predict comparable kinds of judgments? Many moral judgements are context-sensitive Sinnott-Armstrong (2008). Even if one thinks of an out-group member as having a lower status, they might elevate that status when considering them in comparison to another kind of out-group. By outlining the contours of the relevant context, our findings can help specify the social meaning of partisan dehumanization.

**Hypotheses**

As we described above, prior work reveals the utilitarian option to the theoretically expected default for ‘lever turning’ or ‘switch flipping’ versions of the trolley problem. The utilitarian hypothesis may be formulated:

**H1:** Partisans will be willing to sacrifice a single co-partisan (out-partisan) for the sake of saving five co-partisans (out-partisans).

There is little reason to think that one’s party membership is a relevant fact in the overall utility calculus, or at least not the kind of fact which, in isolation from anything else, could exceed the value of saving four additional lives. This is especially true when the partisanship of the six people (5 on the track and 1 to be sacrificed) is held constant.

However, if polarization influences moral judgement as well as affective responses, then we anticipate that partisanship will also affect responses. First, partisanship would predict that individuals would be less willing to sacrifice members of their own party than members of the opposing party.

**H2:** Partisans will be more willing to save a co-partisan group than opposing partisan group.

**H3:** Partisans will be less willing to sacrifice a co-partisan than an opposing partisan.

If partisans are both less willing to sacrifice their own and more eager to save their own, then then H2 and H3 together anticipate the greatest support for turning a trolley to sacrifice one out-partisan for the sake of five co-partisans. The account expects the least support for sacrificing a co-partisan for the sake of five out-partisans.
Data and Methods

To test these hypotheses, we use an original survey experiment that was embedded in an individual module of the 2018 Cooperative Congressional Election Study (CCES). The CCES is a large-n survey in which teams contribute individual modules of approximately 1,000 respondents \((N \approx 60,000)\).

Respondents were introduced to the trolley problem with the following sentence, “In the next four questions we will ask you to make a series of ethical judgements.” Respondents were then presented the traditional trolley problem and asked if they would flip the switch to save the five stranded individuals by sacrificing a single person. However, in each round of the four, the identity of the five potential victims and the sad individual who would be sacrificed to save the five was randomly assigned. The five individuals who would be saved by flipping the switch were randomized to be either five Republicans or five Democrats. The individual who would be offered as a sacrifice to save the five people was also chosen at random from seven possible identities: 1. a Republican, 2. a Democrat, 3. a convicted criminal, 4. an illegal immigrant, 5. a neo-Nazi, 6. a golden retriever dog, or 7. a non-US citizen.

While respondents were asked to complete four scenarios with different characteristics randomly chosen in each round, some chose not to and skipped questions leading to a small amount of item non-response. Together this provides us with 1,597 different responses to random combinations of five partisans being potentially saved by sacrificing one of seven possible victims. This is a much larger sample of responses than previous work testing variations of the trolley problem in other domains and is a dramatic improvement in providing precise estimates of any observed effects.

Our key interest is in how the partisan identity of the five individuals on the track, combined with the partisan identity of the respondent and the identity of the individual being sacrificed to save the five impacts the likelihood of making the decision to flip the switch and offer the individual sacrifice to save the five partisans. Our hypotheses, described above, suggest that partisanship (whether shared or not between the respondent and the group or the individual sacrificed) should play a meaningful role in this decision. We test this using a simple linear regression in which the dependent variable is 1 if the respondent chose to save the five individuals by sacrificing the single person (or dog in one of the seven cases). As independent variables we include the identity of the sacrificial individual while also considering the partisanship of the five people and the partisan affiliation of the respondent.
Results

As the utilitarian hypothesis (H1) would predict, there is a high absolute level of support for turning the trolley to save five in all cases. However, as the partisanship hypotheses predict, party identity plays a significant role in diminishing one’s willingness to turn the trolley. Figure 1 displays the proportion of respondents who would turn the trolley in the case when the five individuals to be saved are co-partisans and when the five are out-partisans. It also shows the proportion pulling the lever when the one person to be sacrificed is a co-partisan or out-partisan.

The cases in which the respondent is saving five co-partisans (left panel) align with the utilitarian hypothesis. Regardless of whether the one individual being sacrificed is a co-partisan (.90) or out-partisan (.92), respondents are very likely to pull the lever to save five co-partisans and the difference in the partisan identity whether the individual being sacrificed has no meaningful impact.

However, things look very different when the five individuals being saved are members of the out party (right panel). In these cases, there is strong support for the partisanship hypotheses, H2 and H3. H2 suggested that partisans will be more willing to save groups of co-partisans than opposing partisans. Figure 1 shows strong evidence of this relationship. Comparing the first two bars to the third and fourth bars shows exactly this. When rescuing a group of co-partisans, respondents are willing to sacrifice an out-partisan 92% of the time while respondents were only willing to sacrifice an out-partisan 75% of the time to save a group of five out-partisans (diff = .18, p < .001). Evidence for H3 is mixed. Republicans and Democrats alike are less willing to sacrifice a co-partisan than an opposing partisan for five opposing partisans, but will sacrifice both a co-partisan and an opposing partisan at a high right for five co-partisans.

As expected by H2 and H3 together, there was least support for sacrificing a co-partisan for the sake of an opposing partisan group. When saving five co-partisans, respondents are willing to sacrifice a co-partisan 90% of the time while respondents were only willing to sacrifice a co-partisan 59% of the time to save a group of five out-partisans (diff = .31, p < .001).

What might account for the effect of partisanship on willingness to turn the trolley? One possibility is that respondents regard it as morally unacceptable to sacrifice a co-partisan in circumstances in which they would regard it as morally acceptable to sacrifice an out-partisan. If it is wrong to treat one agent in some way where it is not wrong to treat other agents, this is – ceteris paribus – a way of expressing the idea that the former agent has a higher moral status. According to what we might call the moral status explanation: It is wrong to sacrifice co-partisans in some circumstances in which it would not be wrong to sacrifice opposing partisans.
Moral philosophers have traditionally identified personhood with relatively high moral status Darwall (2006). If an agent is a person, a greater variety of actions toward that agent may be objectionable, and so the agent has higher moral status. Persons not only experience pleasure and pain states, but are also capable of reflectively organizing their experiences into a narrative self-conception. Because persons can conceive of themselves as temporally extended agents, others may be morally required to likewise regard them as holding moral claims over their complete lives. According to the “separateness of persons” thesis, it is therefore wrong to impose some burdens on individual persons for the sake of greater overall well-being, even though it would not be wrong to impose similar burdens on beings with lower moral status Otsuka and Voorhoeve (2009).

The moral status explanation coheres well with the partisan dehumanization finding. Political dehumanization involves viewing opposing partisans as less advanced agents Martherus et al. (2019); Cassese (2019). Research in non-political domains has found that dehumanized persons are more easily subject to otherwise objectionable forms of treatment. If co-partisans award each other a higher moral status than out-partisans, then we would expect them to be more reluctant to sacrifice fellow in-group members. Our findings do not support the interpretation that agents are never willing to sacrifice others with high moral status, as co-partisans will sacrifice one of their own to save five of their own at a very high rate (90%). They might yet
regard it as objectionable to sacrifice a higher status agent for a group of lower status agents.

Other interpretations are possible. One alternative to the moral status explanation is what we might call the partiality explanation: *It is permissible to withhold from sacrificing a co-partisan in some circumstances in which it would not be permissible to withhold from sacrificing opposing partisans.*

Partiality is distinct from moral status. According to philosophical explanations of partiality, it is the relationship one has to another agent, rather than that agent’s properties, that creates a moral permission (Klodny, 2010). For example, one might be justified in refusing to sacrifice a loved one to save others, even if they would be constrained to sacrifice one stranger to save five. Such a view does not imply that their loved ones differ in moral status from anyone else (Setiya, 2014). With respect to the political science literature, partiality is associated with warm feelings toward co-partisans, whereas negative partisanship, with its associated dehumanization, would predict treating out-partisans as occupying a lower moral status.

A third interpretation is that partisans simply want to maximize the number of their fellow partisans. We might all this political consequentialism: *It is morally acceptable to sacrifice a co-partisan or an out-partisan if it maximize the number of co-partisan lives saved.* Although our vignettes gave no indication that increasing the total or relative numbers of co-partisans would serve some greater political end, it’s possible that partisans independently attach instrumental value to bolstering their own ranks. Although we find high levels of absolute support for saving opposing partisan groups, there is less support than for co-partisan groups.

All three explanations can account for the preference to saving co-partisans to out-partisans. Only the moral status explanation requires a negative attitude toward out-partisans, as it alone suggests that they are vulnerable to forms of treatment from which in-group members should be protected. If the moral status explanation accounted for the success of the partisanship hypotheses, we would expect that partisans would not only prefer co-partisans, but that they would also regard out-partisans as comparable to groups with lower moral status. Members of some groups might be regarded as deserving less moral protection for a variety of reasons. Some (e.g. noncitizens, illegal immigrants) may be perceived as lacking a legal basis for some moral protections (Utych, 2018); Louis, Esses and Lalonde (2013). Others (e.g. convicted criminals) may be thought to have alienated their protections against punishment (Bastian, Denson and Haslam, 2013). Others (e.g. neo-Nazis) represent views that deprive other persons of moral status. Finally, members of some groups are not persons at all (e.g. dogs).

Figure 2 shows the proportion of “switch flipping” to save 5 co-partisans and 5 out-partisans. Among all respondents, a sacrificial out-partisan is the most likely to trigger a ‘flip’ when saving 5 co-partisans, and is
Figure 2: Trolley Problem, Partisans versus Other Groups

Note: Rates of sacrificing 1 and save 5 among seven different sacrificial individuals. Stars indicating statistical significance (p < .05) are in reference to sacrificing an out-partisan within each panel. The top panel shows rates of sacrificing to save 5 co-partisans. The bottom panel shows rates of sacrificing to save 5 out-partisans. Hash marks indicate 95% confidence intervals.

second only to neo-Nazis when saving 5 out-partisans. These results suggest that among other candidates for lower moral status, out-partisans sit quite comfortably among neo-Nazis, criminals, and dogs.

Figure 3 shows whether these preferences differ by the party of the respondent pulling the hypothetical lever. The top panels shows rates of sacrificing 1 to save 5 co-partisans. The bottom panels shows the rates sacrificing 1 to save 5 out-partisans. Republican respondents are shown on the left panels and Democratic respondents are shown on the right panels. For Republicans, we find no statistical difference in how they respond to saving a group of fellow Republicans (top left panel). When saving a group of Democrats (bottom left panel), Republicans exhibit a strong aversion to sacrificing one of their own, but otherwise show no statistical difference among other sacrificial candidates. In other words, Republicans display no partiality toward Democrats, even relative to neo-Nazis, convicted criminals, or golden retrievers.

Democrats, when saving 5 fellow Democrats (top right panel), treat Republicans on a par with neo-Nazis, convicted criminals, and surprisingly, fellow Democrats. They are less willing to sacrifice an immigrant, non-citizen or dog to save 5 fellow Democrats. When saving 5 Republicans (bottom right panel), Democrats show more discrimination across the seven different options. Neo-Nazis are most likely to be sacrificed, fol-
lowed closely by a Republican. A fellow Democrat as the sacrifice to save 5 Republicans is the least likely outcome. In short, of those groups considered, Republicans care only about shared Republican identity. Democrats might as well be criminals or Nazis. For Democrats, Republicans are slightly favored to Nazis, but only to Nazis.

**Discussion**

Our findings offer evidence that partisan loyalties do extend to moral judgments. Negative partisan attitudes appear reactive – directed toward opposing partisans themselves, rather than merely targeting circumstances of inter-partisan interaction. Finally, these attitudes appear quite serious. People treat out-partisans comparably to other dehumanized and denigrated groups. Partiality to co-partisans cannot explain the comparison between out-partisans and the most extreme outgroups we considered. Congruent with other findings affirming the pervasiveness of negative partisanship, our results appear driven at least in part by negative attitudes toward political opponents. In our case, these negative attitudes include not only affect, but the judgment (at least, the expressed judgment) that out-partisans occupy a lower moral status.

At the outset, we noted a kind of normative aspiration to civic friendship as an ideal of shared citizenship. Our results tend toward pessimism about this normative ideal. There is little indication that partisans invest much positive value in shared citizenship. The idea that co-citizens, even of opposing political tribes, share a common project of ruling together, and further that this common project gives them special obligations to each other, is absent from our picture Scheffler (2010); Kolodny (2014). Insofar as they require that opposing partisans share a valuing relationship Scheffler (2005); Rawls (2005); Viehoff (2014), normative theories of citizenship look untethered from political reality.

However, other normative theorists affirm a distinctive normative value to partisan attachment. These theorists see partisanship as an expression of a political commitment that makes ongoing political action possible Ypi (2016). Our results offer grounds for a more sanguine perspective on this value. However, our findings also offer a cautionary note for proponents of partisan loyalty. Such bonds appear not to be constituted merely by partiality to one’s political allies. They include, as well, a willingness to compare opponents with disliked and even reviled groups. This may extend to seeing them as less deserving of moral concern. The partisan ideal may be one about which one might be appropriately cautious – and not only when approaching a trolley crossing.
Figure 3: Trolley Problem, Partisans Versus Other Groups, by Party of Respondent

Note: Rates of sacrificing 1 and save 5 among seven different sacrificial individuals by party of survey respondent. Stars indicating statistical significance ($p < .05$) are in reference to sacrificing an out-partisan within each panel. The top panel shows rates of sacrificing to save 5 co-partisans. The bottom panel shows rates of sacrificing to save 5 out-partisans. Hash marks indicate 95% confidence intervals.
References


Vallier, Kevin. 2019. Must politics be war?: Restoring our trust in the open society. Oxford University Press, USA.


