Mandatory Arrest vs. Arrest with Warrant, Domestic Violence and False Accusation

Parimal Bag[*] Brishti Guha[†]

September 12, 2025

Abstract: For violent men and malicious women in marriage and cohabitation (not necessarily paired together), battering and false accusation are strategic complements. This study compares the implications of laws that require a warrant for arrest vs. mandatory arrest for violence in the presence of malicious reporting. Making arrests based on women's reports, even in the absence of substantive evidence, clogs trial courts with false malicious cases, a noise that enables violent men to batter more. The prior requirement of an arrest warrant places a heavier evidentiary burden on the police before cases come to the prosecutor, together with the need for costly production of fake injuries by malicious plaintiffs. We find that under warrant-based arrests, false accusations can be ruled out on a wide range of violence levels, and real perpetrators can be convicted with minimal burden on judicial resources. Using appropriate (plea) penalties, severe violence is easiest to deter under warrant mandates, while it is hardest to deter under mandatory arrests. Warrants also help bring charges proportionate to the severity of crimes. Thus, we shed new light on the merits of the mandatory arrest law for DV.

JEL Classification: K41, K42, C72, D71, D82.

Keywords: Battering, false accusation, warrantless arrests, arrest with warrant, plea bargains, trials, costly juror effort.

[*] Department of Economics, National University of Singapore, Faculty of Arts and Social Sciences, AS2 Level 6, 1 Arts Link, Singapore 117570; E-mail: ecsbpk@nus.edu.sg

[†] Centre for the Study of World Economy, School of International Studies, Jawaharlal Nehru University, New Delhi 110067. Email: brishtiguha@gmail.com

1 Introduction

Three-fourths of the violence experienced by women in their lifetimes is perpetrated by their partners (Aizer, 2010). The economics literature on intimate partner violence has focused on the problem of domestic violence (DV), studying various possible causes.[1] Parallelly, legal scholars and practitioners, as well as non-economics literature on DV, have emphasized the problem of false reports by potential victims. The latter issue, to our knowledge, has not been modeled in the economics literature. Incorporating false accusation in reporting, we study what kind of case admission rule—mandatory arrest (conditional on mere reporting of violence) or a more demanding warrant-mandated arrest—is better for prevention of DV.

Rutledge (2009) documents that false statements in DV cases were so frequent as to be "considered an epidemic with 80-90% victims lying." This includes both false allegations and true ones where the victims later wanted to recant their statements because of fear, economic considerations or due to relationship-induced complications. He notes that after 1980, warrantless arrests for misdemeanor assaults were implemented, and mandatory arrest policies were implemented by 47 states in the US allowing police officers to arrest purported offenders on the basis of a report without any corroborating evidence; see also Zeoli, Norris and Brenner (2011). While intended to protect real victims, this also made it easy to lodge false reports.[2] Avieli (2022) documents the wide prevalence of false allegations of DV, noting that some studies estimated false allegations as constituting up to 50% of all allegations. Estimates in specific cases have been even larger; for instance, Foster (2007) estimates, using information on court decisions in West Virginia, that 80% of DV petitions filed were either false or unnecessary. Allegedly false claims of DV have also been hitting headline news in different countries recently.[3]

[1] For instance, Tauchen et al. (1991) discuss how some men may experience gratification from using violence to control their wives. Dowry violence in arranged marriages is a major concern in India (Bloch and Rao, 2002). Aizer (2010) and Anderberg et al. (2016) find empirically that a reduction in the gender-wage gap reduces DV, while Arenas-Arroyo et al. (2021) find that an increase in women's relative income worsens the violence they experience from partners, due to "male backlash". Some studies, like Tur-Prats (2021), find that whether the male backlash effect is strong or not depends on cultural factors. Studying individual level data on DV in India, Anderson and Genicot (2015) found that increased property rights for women led to an increase in wife beating.

[2] Similar changes were happening in other countries; for instance, in India, husbands accused of violence under section 498A of the Indian Penal Code could be arrested without warrants, and these cases could not be taken back by the complainant, mirroring no-drop prosecution laws in the US.

[3] One such high-profile case involved actor Johnny Depp and his ex-wife, Amber Heard, accusing him of violence, resulting in several counter-suits of defamation, with Depp receiving compensation of $15 million in Depp vs Heard 2022. The jury in the case decided that Heard's claims of violence were false and motivated by malice. Another ongoing case in India is the Atul Subhash case, where a husband committed suicide, leaving behind a 24-page suicide note and a 80-minute video imputing his suicide to false legal allegations of violence and extortion lodged against him by his wife and her relatives. Concerns over false allegations of DV have led to a public interest litigation being filed in India, asking the Supreme Court to reform loopholes in the laws meant to protect women from DV, as these loopholes allowed the misuse of these laws in false cases.

As pointed out in the public interest litigation cases mentioned above (footnote 3), the reality of false allegations has consequences for the many true incidents of DV, creating an atmosphere where “the real and true incidents against women are looked at with suspicion,” (in the petitioner’s words). Thus, the problem of false allegations does not in any way detract from the very real problem of DV; in fact, it makes it harder for real victims, and in addition, creates another category of victims – innocent men.

In the current paper, we tie up themes of DV with (i) the possibility of false accusations, and (ii) plea bargaining. In spite of the extensive use of plea bargaining in dispute resolution (in the US, 95% of criminal cases are settled by plea bargaining; see Berg and Kim, 2018), and despite there being a substantial literature on plea bargaining in law and economics (Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezetti, 2001; Kim, 2010; Lee, 2014; Bjerk, 2007, 2021; Guha, 2023, 2024, and many others), no paper has, to our knowledge, explored plea bargaining as a tool to (a) deter DV, and (b) discourage false reports of DV. We explore these questions. Due to the near-ubiquity of plea bargaining in many countries and its growing popularity in others (Turner, 2013), it is important to examine the twin problems of genuine DV and false accusations of DV afresh. In addition, plea bargaining is a more flexible tool than changing penalties, as legal reform is generally not a swift process.

■ **Results: Our contribution to a debate.** There is an interesting debate in the literature on the effectiveness of warrantless, mandatory arrest policies for DV. Proponents believe that mandatory arrest would serve as a powerful deterrent to DV. By ensuring that police arrest abusers when a DV call is made, such laws would, in theory, ensure that perpetrators were punished more often (keeping reporting constant); at the same time, they might also encourage reporting as victims could be sure police would respond to their reports of DV. However, the empirical literature has found contrary evidence. Iyengar (2009), for example, finds that mandatory arrest actually increased intimate partner homicides. She attributes this to the possibility that mandatory arrests may reduce the victim’s willingness to report, as she may not want her abuser to be arrested; thus, giving the abuser an opportunity to commit homicide. Later, Chin and Cunningham (2019) find that while mandatory arrest laws do not increase intimate partner homicides, discretionary arrest laws (where the police officers had the discretion on whether to make a DV arrest) decrease such homicides. Thus, it is not clear whether mandatory arrests work better than policies with a ceteris paribus lower probability of arrest.

We shed new light on this debate by incorporating the possibility of false, malicious accusations of DV. Moreover, we account for the fact that potentially violent men can differ in the magnitude of harm they inflict; they may be heterogeneous in their propen-

See https://en.wikipedia.org/wiki/Depp_v._Heard and https://www.ndtv.com/india-news/plea-in-supreme-court-calls-for-reform-in-dowry-and-domestic-violence-laws-7238508.

sities for violence. We then contrast a mandatory arrest policy with a policy where DV arrests require a warrant. More generally, the second policy requires officers to investigate; if officers fail to secure a signal that the complaint is genuine, they cannot proceed. If the complaint is genuine, and the officers secure a signal, their investigation also reveals the level of violence involved—evidence that is passed on to the prosecutor and the jury (in the event of a trial).[4] Women who file malicious reports would now need to fake evidence (since an investigation will otherwise uncover that there has been no violence), an exercise that is costly, particularly if the woman accuses the man of a high level of violence. Thus, while under mandatory arrests, all true reports of battery enter the system, so do all false reports. The second, evidence-based policy does introduce a possibility that some true cases, even if reported, will not be followed up on. However, at the same time, it also makes it more difficult for false reports to enter the system owing to the fact that deceiving the investigators is costly.

Interestingly, we find that some levels of violence will never be faked. Hence, if the prosecutor or the jury receive a report of such a level of violence, they correctly believe that the defendant is guilty. Then, conviction in the event of a trial is certain, and can be done without costly judicial expenditure. This also allows plea bargains to be harsh, as the prosecutor does not have to include discounts due to the possibility of false cases, or judicial errors. Moreover, the victims of DV realize that if their case enters the system, they can definitely secure a conviction, a fact that encourages reporting.

More violent crimes are actually easier to deter under this policy. Now, charges can be brought for the specific level of violence uncovered by the police investigation. We find that penalties and plea offers for more severe charges need to satisfy a relatively modest threshold for them to deter the violence levels that result in these severe charges.[5] This is in sharp contrast to a mandatory arrest policy, where it is easier to deter intermediate levels of violence, and where the more violent crimes are the most difficult to deter. (Because mandatory arrest does not involve information from detailed investigations, penalties cannot be tailored to the actual level of violence. Hence, the plea bargain is not severe enough to deter highly violent men, while it might deter men with a smaller propensity for harm. Crimes involving very low violence might not be reported, as women are unsure of obtaining a conviction.)

Thus, deterrence, particularly for more serious crimes, can actually worsen under

[4] Investigation entails a fixed cost for the police. Hence, investigation is undertaken only when the law requires it (i.e. arrests require a warrant). Thus, similar investigations are not undertaken when arrests are mandatory and warrantless. Although we look at two polar situations, arrest with a warrant and mandatory warrantless arrests, it is easy to use our analysis to draw inferences about intermediate regimes, such as discretionary arrests.

[5] An intuition for this comes from the investigation technology. More violent crimes are disproportionately more likely to result in a signal, so that they are entered into the system; the probability of being detected by the police increases faster than the level of violence. This seems realistic; Zeoli et al. (2011) mention how more severe violence also produces more evidence of violence.

mandatory warrantless arrest laws. Moreover, the mandatory arrest regime involves more expenditure of judicial resources and no deterrence of false, malicious reporting. False reporting, however, is deterred for a wide range of violence levels under evidence- or warrant-based arrests. This, in turn, helps deter actual violence as well, as explained above.

Therefore, our results confirm that mandatory arrest policies have nuanced implications. Ours is the first paper of which we are aware that theoretically approaches this problem. In doing so, we have also stressed a novel channel in the literature – the possibility of some accusations being maliciously motivated, and the strategic complementarity between true violence and false accusations. (By strategic complementarity, we mean that the prevalence of false accusations can enable real perpetrators to hide behind them. False reports, in turn, are only made possible by the fact that real cases are also happening; otherwise, these reports would be marked out as false.) Different arrest laws affect this channel differently, driving our results and providing an explanation for the observed empirical results which is distinct from possible explanations highlighted by the empirical literature. In addition, we have derived interesting implications about the relative severity of the crimes that can be deterred under different legal regimes, and about the relative burdens on judicial resources in warrant or evidence-base regimes versus mandatory, warrantless ones.

■ **Literature review.** Our paper connects the literature on DV outlined above to the law and economics literature on plea bargaining and costly juror effort. For modeling of violence and reporting, we draw upon Aizer and Dalbo (2009), although the core issue studied by the authors is quite different. They look at how changes in (woman) victim's emotions, from withdrawal and anger right after battering by her man to a more softening and forgiving posture with a time lapse, can cause her to drop the charge. A no-drop policy would save the victim from this time-inconsistency problem, leading to greater reporting and lowering of the probability that the battering man might be killed (an extreme form of commitment to avoid future battering in the absence of a no-drop policy). Our model instead assumes rational weighting of the costs and benefits of reporting, and there is no extreme option for the woman to kill her batterer. Our main focus is on how to separate the true accusation of violence from the false reporting.

The two papers that are much closer to the current paper in terms of policy relevance, Iyengar (2009) and Chin and Cunningham (2019), have already been discussed above.

The problem of false accusations of abuse in social interaction settings was also studied by Lee and Suen (2020). The authors consider a two-period model involving a potential abuser (say, a person prone to sexual misconduct) and multiple potential victims. They show that real victims have an incentive to immediately report abuse. This is because such victims believe that the abuser may have had other victims in the past (who would be inspired to submit a delayed report of abuse, if the new victim reports now) or may abuse someone else in the future, who would also be encouraged to report. However, those who make false accusations know that the probability of past or future accusations that corroborate theirs is low. Thus, even if true and false accusers have the same payoff, they behave differently in what the authors call a "corroboration equilibrium". The paper does not explicitly model the justice system (either jury trials, or plea bargaining). The similarity to our paper is that a framework which achieves semi-separation in a context with both true and false accusations is considered. We differ from their paper in abstracting from multiple periods, in modeling plea bargaining and jury trials (which helps us explore the role of instruments such as penalties and plea bargaining), and in focusing on the interaction of true crime and false accusations with the arrest laws – mandatory, warrantless arrests versus evidence and warrant-based ones.

Mukhopadhyaya (2003) studied the problem of jurors paying attention to trial-relevant information and had shown that due to costly juror effort sometimes a small jury panel can make more accurate decisions than larger ones. In a number of papers, Guha (2018; 2020; 2023; 2024) studied costly juror efforts and plea bargaining, observing that (i) small effort costs can lead to free riding and worse verdicts even as priors get more informative, (ii) often there is no pooling equilibrium where all defendants accept the same plea offer (reducing common concerns about false guilty pleas by the innocent), (iii) in accomplice plea bargains, a prosecutor wanting the guilty to be harshly punished and the innocent to be spared, can achieve perfect sorting; etc.

Costly juror effort also has a basis in the empirical psychology literature. Bornstein and Greene (2011) discuss the cognitive limits of jurors, finding that jurors resort to heuristics and are subject to biases such as availability bias (which relates to the ease of recalling information), hindsight bias and representativeness bias. Overcoming ingrained psychological biases is taxing, providing a reason why jurors might find it costly to pay attention during a trial.

Other papers on plea bargaining that endogenized juror decision-making ignore costly juror effort, e.g., Lee (2014) and Bjerk (2007, 2021). In Lee, prosecutors who care more about wrongful conviction than wrongful acquittal can reduce the conviction rate by using plea bargaining. In Bjerk, both the prosecutors and the jury receive signals of guilt, with the jury receiving more precise signals, and prosecutors choosing a threshold signal above which they deny plea bargains and send everyone to trial. None of these papers deals with domestic violence or false reports.[6]

In Siegel and Strulovici (2023), plea bargaining emerges endogenously as a feature of judicial mechanism design. Our paper also relates to Siegel and Strulovici (2025),

[6] Berg and Kim (2018) show that accomplice plea bargaining reduces crime as accomplices spread information about criminals. They do not consider single-defendant plea bargaining, so their entire effect comes from the reports spread by accomplices.

which discusses how welfare gains may be achieved by penalties that are graduated to the strength of evidence; rather than simply having binary punishments – a uniform penalty for guilt, and zero penalty for acquittal – they discuss intermediate penalties for lesser evidence. They also point out that the incentives for a third party, such as police investigators, to gather presumably costly evidence are minimal if penalties are not in fact evidence-based. These relate to our paper, because in our analysis of laws requiring warrant-based arrests, we allow penalties if convicted at trial to be sensitive to the evidence uncovered by police investigators. We also discuss how incentives for police to do similar investigations disappear under mandatory, warrantless arrests where penalties are uniform.

Finally, the paper draws on Becker's seminal insights on crime and punishment (Becker, 1968; 1993). Both conviction and plea bargain penalties must be large enough to minimize violence, while given the likely errors in verdicts, penalties can also be set minimally. Guided by this basic principle, we focus on how to best use the instrument of arrest laws to maximize the probability of guilt in the pool of defendants (conditional on keeping crime incentives in check) so that jurors are diligent in their deliberation to be able to deliver correct verdicts.

The remainder of the paper is organized as follows. Section 2 sets up the basic model. Section 3 analyzes a benchmark case in which all violent men are homogeneous and perpetrate the same degree of harm. Our main results are in Sections 4 and 5. Both deal with the case where violent men are heterogeneous and may perpetrate different degrees of harm; while Section 4 deals with mandatory, warrantless arrests, Section 5 deals with a regime where arrests require evidence, or a warrant. Section 6 concludes. An appendix contains the proofs, while a supplementary appendix reports some extra results.

2 Model

There are two representative partners (a man, M and a woman, W), an uninformed prosecutor, and jurors. The man may be intrinsically non-violent with a proportion α of men of this type, or a violent type who derives the utility $v > 0$ from violence that we call *ego value*, v being common knowledge.[7] However, male partner's type is unknown to outsiders such as the prosecutor or jurors, although it becomes known to the woman during day-to-day interactions.

A fraction s of women, even when living with non-violent men, are prepared to falsely accuse them of violence tempted by potential gains that we simply call *malice*.[8] Let $s < 1$

[7] This utility is akin to gratification from male dominance as in Tauchen et al. (1991). A small trigger in the form of difference of opinion can prompt such behavior.

[8] Non-violence doesn't necessarily mean a happy marriage, and a functioning relationship can give rise to malice.

as at least some women find it morally distasteful to lodge false reports.

Given α and s that are exogenous, a randomly drawn (M, W) pair can be one of four types with associated probabilities: (violent,non-malice; $\alpha(1-s)$), (violent,malice; αs), (non-violent,malice; $(1-\alpha)s$), (non-violent,non-malice; $(1-\alpha)(1-s)$). Whenever a man is violent, the malice type is of little concern because ideally both malice and non-malice types should report; worries of false accusation do not apply. In the current paper, we do not consider any causal link between malice and violence.

Violent types must decide between battering or not the women they live with. Denote this action by $a \in \{B, NB\}$. Some men, whether they are actually violent or not, may be reported by their partners. Let $\chi_\theta \in \{R, N\}$ denote the reporting strategy of the woman, where $\theta = B$ means that she has been battered and $\theta = NB$ means not battered, and R means a report of battering and N means no such report. Sometimes we will use superscript m or nm to χ_θ to clearly indicate the strategy of a malicious or non-malicious woman. Following a report of battering, the man has to appear before a prosecutor. The prosecutor then offers a plea bargain, a penalty p_L, and the defendant decides whether to accept or reject it.

Inflicting of violence and subsequent decisions proceed as in Fig. 1. Fig. 2 is a complete description of the extensive form interaction between any two partners with the payoffs indicated at terminal nodes. This latter game tree is an adaptation of Fig. 1 of Aizer and Dalbo (2009) for our problem.[9] A woman may report a man to the authority (police) for battering, truthfully or falsely, following which the prosecutor offers plea bargain (of reduced punishment) to the accused perpetrator. If the perpetrator declines the offer, the case goes to trial and jurors deliberate by paying attention or not to any evidence of battering brought before them. Finally, the jury may convict or acquit the defendant. Acquittal may be unqualified or comes with the additional message that the plaintiff was malicious in her intent (in short m.i.).

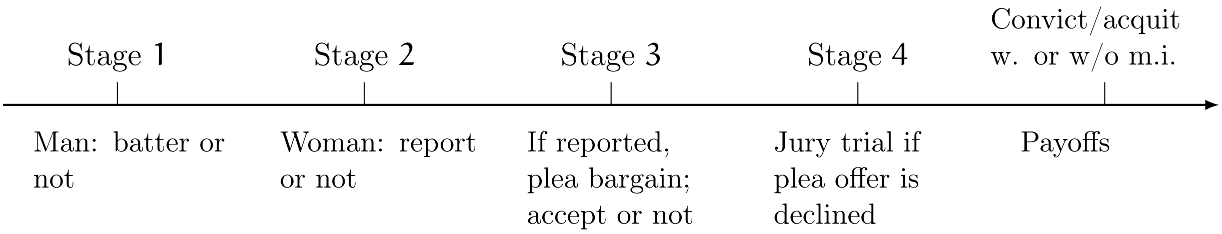

Stage 1
Stage 2
Stage 3
Stage 4
Convict/acquit w. or w/o m.i.
Man: batter or not
Woman: report or not
If reported, plea bargain; accept or not
Jury trial if plea offer is declined
Payoffs

Figure 1: Time line

The first two stages of the game (see Fig. 1) are self-explanatory. Let us describe the prosecutor's problem. The prosecutor faces a defendant, a man accused of battering his

[9] In Aizer and Dalbo, the victim had the following options that are different from ours: (i) victim could not lodge a false report, (ii) victim could kill the battering partner, and (iii) victim could drop a case. Our victim may lodge a false report, cannot kill a partner, and once a report for battering is made the complaint cannot be withdrawn. In addition, we have added a plea bargain phase and jury trial.

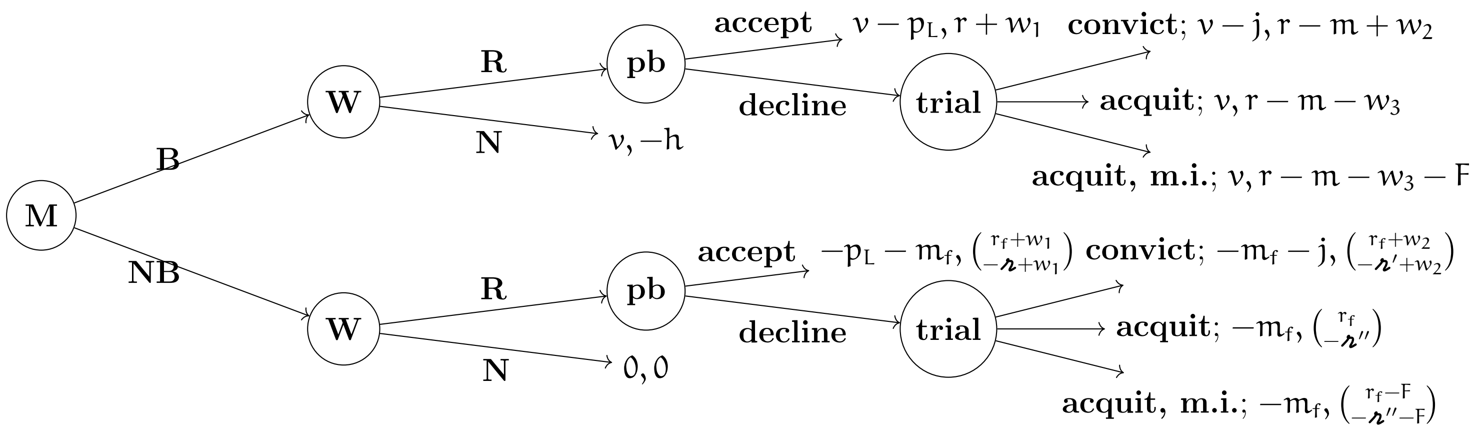

M
B
NB
W
R
N
pb
accept
decline
trial
$\nu - p_L, r + w_1$
convict; $\nu - j, r - m + w_2$
acquit; $\nu, r - m - w_3$
acquit, m.i.; $\nu, r - m - w_3 - F$
$\nu, -h$
$-p_L - m_f, \binom{r_f + w_1}{-\varkappa + w_1}$
convict; $-m_f - j, \binom{r_f + w_2}{-\varkappa' + w_2}$
acquit; $-m_f, \binom{r_f}{-\varkappa''}$
acquit, m.i.; $-m_f, \binom{r_f - F}{-\varkappa'' - F}$
$0, 0$

Figure 2: This game is both a simplification and an extension of the game of domestic violence in Aizer and Dalbo (2009); see Fig. 1 in their article. Explanation of notations/payoff parameters: **m.i** means malicious intent by the plaintiff; p_L=lighter punishment; ν=utility from battering; j=loss from conviction (e.g., jail time) where $j > p_L$; m_f=loss to the defendant from malicious trial; h=mental scar/harm on the woman from battering; r=utility to woman from truthful reporting; m=loss to true victim (woman) from trial stress; w_1=satisfaction plus compensation from successful plea verdict ($p_L > w_1$, if w_1 is solely the compensation paid by defendant); w_2=satisfaction plus compensation from conviction, so $j > w_2$ (if w_2 is solely the compensation from defendant) and $w_2 > w_1$; w_3=mental scar plus anger from wrong verdict (so $w_3 > h$); F=penalty for false accusation of violence; r_f=malicious enjoyment from false accusation; r=feel-good enjoyment of doing the right thing by reporting battering, same for both malicious and non-malicious women; $-\varkappa$=disutility to non-malicious women from false accusation when the defendant accepts plea bargain; $-\varkappa'$=disutility to non-malicious women from false accusation when the verdict is to convict; $-\varkappa''$=disutility to non-malicious women from false accusation when the verdict is to acquit; we impose a natural ordering: $\varkappa' > \varkappa > \varkappa'' > 0$. Finally, the symbol $\binom{x}{y}$ is the listing of payoffs of the two types of plaintiff, the top one corresponding to malicious type and the bottom one for the non-malicious type.

partner. She knows that the defendant may be either guilty, or innocent (in case of a false report), but does not know which. She assigns a probability of guilt, μ, which is the probability that any man accused of violence actually is guilty. The probability μ can be expressed in terms of parameters α and s as follows. Suppose that all men with a propensity for violence chose to be violent, and were reported by their partners, and suppose that all malicious women who can profit by making false reports choose do so. Then we would have

$$\mu = \frac{1 - \alpha}{\alpha s + 1 - \alpha}, \tag{1}$$

assuming random matching between men and women. Here, the numerator of the RHS of (1) indicates the mass of intrinsically violent men, while the denominator contains them, as well as the intrinsically non-violent men falsely accused by their partners. Note that all violent men are homogeneous (thus if one decides to be violent, all do the same) and all women who lodge false accusations are also homogeneous (with a similar implication). We will relax this last assumption in Sections 4 and 5.

Given this prior probability of guilt, μ, the prosecutor then offers the defendant a plea bargain with (endogenous) penalty p_L. The plea offer will depend on the prosecutor's expectations of how this offer will affect juror's incentives to pay attention to trials and conviction probabilities. If the plea is accepted by the defendant, the game ends with the man's and woman's payoffs indicated in Fig. 2 (explanations given below in the caption). If the defendant rejects the plea offer, he goes to trial, where a panel of n jurors hears the evidence against him, deliberates and decides on conviction versus acquittal, and the payoffs are indicated at the corresponding nodes.

Each juror obtains a utility of 1 if the panel as a whole makes a correct verdict; utility of a wrong verdict is normalized to 0. However, each juror must also decide whether to incur a small cost c during the trial, which is the cost of paying attention to trial-relevant information. Each juror who does incur the cost obtains a private, conditionally independent signal of accuracy q (where $1 > q > \max\{\mu, 1-\mu\}$) of the defendant's guilt or innocence (i.e., the signal is correct with probability q and wrong with probability $1-q$). If more than one juror pays attention, they may receive different signals.

Those who do not incur c do not receive any signal. As each juror wants the correct verdict to be reached, he shares his signal truthfully with the rest of the panel. The other jurors realize this. Since they also want the correct verdict to be reached and since different attentive jurors may receive different signals, they vote according to the signal (guilty or innocent) received by the *majority* of attentive jurors.

In the case where equal numbers of attentive jurors have received guilty and innocent signals, or where no one has paid attention, jurors vote according to their (common) updated belief about defendant's guilt or innocence. While the probability of guilt in the pool of defendants is μ, the juror's belief about a defendant's guilt opting for trial (i.e., μ') is different from μ. To appear for trial, a defendant must have rejected the plea offer, and guilty and innocent defendants will have different probabilities of rejecting a given plea offer. Jurors will estimate μ' through Bayesian updating. Thus, if either no one has paid attention, or there is an equal number of guilty and innocent signals received, the panel assembles and votes by consensus based on μ', convicting if μ' exceeds a threshold (that we take to be 0.5, as is consistent with a preponderance of evidence standard), and acquitting otherwise.

The signal received by an attentive juror is precise relative to the cost of paying attention, specifically,

Assumption 1 *q-$c > 0.5$.*

The assumption is a simplification of the condition that for a juror to pay attention, his (net) expected private benefit of the signal must exceed attention cost, $(q - \max\{\mu', 1 - \mu'\})\,\Pr(\text{juror pivotal}) \geq c$, which implies $q - c > \max\{\mu', 1 - \mu'\}$ where μ' is the juror's updated belief about a defendant's guilt who comes up for trial, and $\max\{\mu', 1 - \mu'\}$ must exceed $1/2$ (standard threshold probability of guilt for conviction).

We focus on a symmetric mixed-strategy equilibrium ($SMSE$), where each juror pays attention with an endogenously determined probability $0 < \sigma < 1$. Although there are pure-strategy equilibria where only one juror pays attention, these equilibria involve a coordination problem (Mukhopadhyaya, 2003). An $SMSE$ avoids these issues and treats all jurors alike, and is also in harmony with the experimental results.[10] Then, denoting p to be the probability of a correct verdict,[11] each juror's utility is given by

$$U_J = p - \sigma c. \tag{2}$$

The first term on the RHS represents expected benefit, recalling a correct verdict yields each juror a utility of 1, while the second represents the expected cost of paying attention.

Accordingly, the timeline of the game is:

1. Intrinsically violent men choose an action a, between B and NB. Non-violent men choose NB.

2. Whether B or NB is chosen, women involved must decide whether to report (R) or not (N). A report always prompts the prosecutor to arrest the accused (warrantless arrest, as in Iyengar (2009)).

3. The prosecutor offers a plea bargain with an endogenous penalty p_L to the defendant. It is rejected by a guilty (violent) man with probability λ_G, and by an innocent (non-violent) man with probability λ_I, where both probabilities are endogenous and functions of p_L.

4. If a plea bargain is accepted in step 3, the game ends. Otherwise, the defendant proceeds to trial, where a panel of n jurors hears the evidence. Each juror independently decides with what intensity to pay attention to trial-relevant information (incurring a cost c) which will give the juror a signal of precision q. After the trial the jurors share information and collectively decide on a verdict. The defendant is convicted or acquitted (with or without an assertion about the plaintiff's malicious intent) and the woman accordingly receives her payoff consisting of satisfaction of truthfully reporting (r) or malice of false reporting (r_f) or guilt of false reporting ($-\boldsymbol{\mathscr{r}}$ if defendant pleads guilty, $-\boldsymbol{\mathscr{r}}'$ if defendant is convicted, and $-\boldsymbol{\mathscr{r}}''$ if defendant

[10] Coordination is a serious problem because courtroom procedures forbid jurors from communicating before the trial. As Palfrey et al. (2017) finds in a threshold public good game experiment, coordination may fail even when group members are allowed to communicate their intent to contribute.

[11] Note the distinction between the probability p and the plea offer penalty p_L.

is acquitted but the plaintiff blamed for false accusation), net of trial stress (m) and any penalty for malicious intent (F), plus any satisfaction or anguish from successful/unsuccessful verdict, w_1 or w_2 or $-w_3$; w_1 and w_2 may also include victim compensation paid by the defendant (court mandated restitution).[12] The defendant also receives payoffs that differ depending on whether he is guilty ($-j$) or not guilty, and whether he has been accused truthfully or maliciously. For more details, see the description in Fig. 2 caption.

The above timeline can now be related back to Fig. 1.

Definition 1 *A perfect Bayesian equilibrium in the four-stages extensive game (in short,* PBE*) depicted in Fig. 2, of violence, reporting and jury trials involving men, women in partnership, the prosecutor and jurors, is a profile of strategies,* $(a, \chi_B, \chi_{NB}, p_L, \lambda_G, \lambda_I, \sigma)$,[13] *and beliefs about the defendant's guilt,* $\mu' = \Pr(\tau = G|\textit{declined plea bargain})$, *such that the following hold when solving the game backwards:*

(i) In stage 4, each juror chooses attention strategy σ *to maximize his expected utility as in a Bayesian Nash game given the strategies* $(a, \chi_B, \chi_{NB}, p_L, \lambda_G, \lambda_I)$ *and beliefs* μ'*. Jurors update their beliefs* μ' *along the equilibrium path using Bayes' rule; beliefs off-the-equilibrium path satisfy Cho-Kreps intuitive criterion.*

(ii) If prosecuted, in stage 3 violent men do not have any incentive to deviate from their equilibrium choice of λ_G*, given* p_L, σ *and beliefs* μ'*, given the stage 4 continuation equilibrium. Likewise, non-violent men have no incentive to deviate from their equilibrium choice of* λ_I*.*

(iii) Fix the equilibrium in the game from stage 3 onward. In stage 2, women who experience battering, B*, choose* χ_B *optimally to maximize their expected payoff in the continuation game given others' strategies and beliefs. Likewise, malicious women who do not experience battering choose* χ_{NB} *(i.e., may lodge false reports) optimally.*

(iv) In stage 1, potentially violent men choose some action $a \in \{B, NB\}$ *to maximize their expected payoff in the overall game. Intrinsically non-violent men choose* $a = NB$*.*

[12] California law on court-mandated restitution to victims of DV covers, for instance, bills for medical treatment and therapy, lost wages etc., and is specified at the time of the defendant's sentence; https://www.sddvattorney.com/blog/101-what-is-victim-restitution-in-domestic-violence-cases. In India, DV victims receive financial compensation both under the DV law and the criminal law; such compensation includes lost wages, treatment for physical injuries, loss due to property such as jewelry being taken away forcibly, and mental torture; see https://nyaaya.org/legal-explainer/how-can-you-get-compensation-or-money-for-domestic-violence/.

[13] Here we write the simpler, although a bit imprecise, notation χ_θ instead of the more accurate strategies χ_θ^m and χ_θ^{nm}.

Jurors' beliefs off the equilibrium path are restricted by the Cho-Kreps intuitive criterion, which specifies that zero weight be placed on the belief that an off-equilibrium deviation has been made by a type that would be worse off after the deviation than in the posited equilibrium.[14]

Throughout our analysis, we assume that the prosecutor "serves the State" in the sense that her main aim is crime deterrence. This is kept in mind when we discuss plea bargains, which are set by the prosecutor; both plea penalties and penalties for conviction (which are exogenously given) are discussed in view of their ability to deter violence, as well as false reporting. In addition, we assume that the jurors will follow majority signals when choosing their verdict without any concern for penalties,[15] and the police will incur any cost of investigation if required by the stipulated standard of evidence for case preparation.

3 Mandatory arrest: Fixed benefit of violence and fixed harm

In this section, we analyze a simpler model, where benefits from violence and harm to the victim are known fixed values. We solve the equilibrium of the four-stage game, where most of the analysis of stages 4 and 3 will carry over to the more general model of heterogeneous violence and harms, analyzed in Sections 4 and 5. This technical development is necessary for the overall paper.

3.1 Stage 4 game fixing prosecutor's plea offer

■ **Juror decision making for a given belief μ'.** Fix any juror belief μ'. To determine σ and p (the probability of a correct verdict), suppose, initially, $\mu' \leq 0.5$ (enough guilty defendants accept the offered plea and stay out of court so that more innocent than guilty men appear for trial). Now, the nth juror is pivotal either if (i) none of the other jurors pay attention, or (ii) if out of the other attentive jurors equal numbers have received opposing signals, in which case there is a tie. In both of these events, if the nth juror is inattentive, the others will acquit, which will be correct with probability $1-\mu'$, while the pivotal juror can ensure a verdict which is correct with probability q if he pays attention. We obtain Eq. (3) below by equating the expected benefit of paying attention to the cost of doing so, which is necessary for him to randomize between paying and not paying

[14] See Cho and Kreps (1987). Suppose a defendant of (action) type $\tau \in \{I, G\}$ (I for innocent and G for guilty) obtains a utility of $u^*(\tau)$ at some posited equilibrium. Then, if jurors observe a defendant deviating from the prescribed equilibrium strategy (say, by sending a message $\widetilde{m}$, rejecting a plea when they are expected to accept it), then if the maximum possible utility type τ derives from such a deviation, given the juror's beliefs, is $\max_{\mu'} u(\tau, \widetilde{m}, \mu')$, then if $u^*(\tau) > \max_{\mu'} u(\tau, \widetilde{m}, \mu')$, jurors believe that a type other than τ made the deviation.

[15] Andreoni (1991) noted that jurors might take into account the severity of penalty when declaring a verdict—letting free a defendant if the penalty imposed is disproportionately high. That is, in Andreoni's world, the penalty and the probability of conviction are not independent.

attention. Thus, we have

$$(q-(1-\mu'))\big[\sum_{j=1}^{(n-1)/2}\frac{(n-1)!}{(n-1-2j)!j!j!}(\sigma q)^j\big(\sigma(1-q)\big)^j(1-\sigma)^{n-1-2j}+(1-\sigma)^{n-1}\big]=c$$

$$\text{i.e.,}\quad (q-(1-\mu'))\big[\sum_{j=1}^{(n-1)/2}\frac{(n-1)!}{(n-1-2j)!j!j!}\big(q(1-q)\big)^j\sigma^{2j}(1-\sigma)^{n-1-2j}+(1-\sigma)^{n-1}\big]=c. \tag{3}$$

Proposition 1 *Suppose* $\mu'\in(1-q+c,\,0.5]$*, and*

$$[q-(1-\mu')]\binom{n-1}{\frac{n-1}{2}}q^{(n-1)/2}(1-q)^{(n-1)/2}<c \quad \textit{for all}\ \mu'\in(1-q+c,\,0.5]. \tag{4}$$

Then the following holds:

(i) There exists at least one symmetric mixed strategy solution σ *to Eq.* (3) *or there will be an odd number of solutions.*

(ii) Consider the case of a unique solution and denote it by σ^**. In the case of multiple solutions, focus on the risk-dominant solution* $\sigma^*=\min\{\sigma_1^*,...,\sigma_k^*\}$ *where* $k>1$ *is an odd integer. The solution* σ^**, be it unique or risk-dominant, is strictly increasing in belief* μ'*.*

(iii) The probability of reaching a correct verdict, p*, is increasing in* μ'*.*

When there are multiple equilibria, the risk-dominant equilibrium is the one with the smallest value of σ. Jurors prefer to play this equilibrium for fear of coordination failure. For σ^* to be positive, μ' must exceed some threshold value because otherwise jurors do not find it worthwhile to incur c.

In the symmetric mixed strategy equilibrium described in Proposition 1, the probability of reaching a correct verdict can be explicitly written as follows (we will write equilibrium σ^* simply as σ):

$$p=\sum_{m=1}^{n}\sum_{k=\lfloor 1+\frac{m}{2}\rfloor}^{m}\frac{n!}{(n-m)!k!(m-k)!}(1-\sigma)^{n-m}(\sigma q)^k\big(\sigma(1-q)\big)^{m-k}$$
$$+(1-\mu')\left[\sum_{j=0}^{(n-1)/2}\frac{n!}{(n-2j)!j!j!}(\sigma q)^j\big(\sigma(1-q)\big)^j(1-\sigma)^{n-2j}\right], \tag{5}$$

where symbol $\lfloor z\rfloor$ denotes the largest integer contained in z. The first term of (5) captures the probability of a correct verdict in the event of no ties. The verdict is correct if the majority of the attentive jurors receive the correct signal. The second term captures ties,

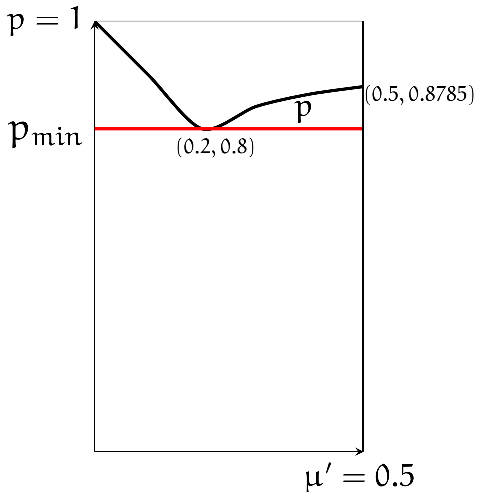


Figure 3: Sample estimates of verdict accuracy p against belief μ'. Numerical calculations are for parameter values $c = 0.1, q = 0.9$ in Table A.3 in the case $n = 3$, supplementary Appendix A.2.

while the case of $j = 0$ covers the possibility that no one is attentive; then an outcome, acquittal, is reached, assuming $\mu' \leq 0.5$, which is only correct if the defendant is actually innocent, that is, with probability $1 - \mu'$.

Note that p is necessarily greater than 0.5. As $1-\mu'$ exceeds 0.5, clearly the verdict is more likely to be correct than wrong even where jurors reach ties, or do not pay attention (if we had assumed $\mu' > 0.5$, the expression $1 - \mu'$ would have been replaced by μ', which would also be greater than 0.5 – in this case the default verdict would have been conviction, which would in the circumstances have been more likely correct than wrong). Wherever they do not reach ties, a majority of attentive jurors are still more likely to be correct than wrong provided signal accuracy is greater than 0.5 (by assumption we have $q > c + 0.5$ so this is always satisfied).

Proposition 2 *Suppose that $\mu' \leq 0.5$. Then the probability of the verdict being accurate, p, attains a minimum at the unique critical value of μ', $\mu'_{cr} = 1-q+c$, with $p_{min} = q-c$. The probability function $p(\mu')$ is of V-shape, as in Fig. 3.*

In Proposition 1 we had established analytically the comparative static effects of beliefs μ' on jury efforts and the verdict accuracy. We did not study the effect of jury size n. In the supplementary Appendix, we do some numerical computations to further illustrate the comparative static effects and demonstrate the effect of jury size n.

3.2 Stage 3 continuation game: Prosecutor's plea offer

We now derive some results. We start by examining the *feasible* ranges of prosecutor plea bargains p_L that would induce continuation games with combinations of $(a, \chi_B, \chi_{NB}, \lambda_G, \lambda_I, \sigma)$ as part of a Bayesian-Nash equilibrium. Note that, as is standard in extensive form games of incomplete information, strategies λ_G and λ_I will also affect beliefs μ', jurors using Bayes' rule to update μ given that a defendant appearing in trial had rejected the plea offer. We start with a lemma.

Lemma 1 *In the plea bargain, guilty defendants are always less likely to reject a given plea offer than innocent ones.*

Proposition 3 *In the plea-bargain stage, neither a fully separating nor a pooling equilibrium exists.*

Thus, any equilibrium where both guilty and innocent types are reported must involve a semi-separating equilibrium, where all innocent types reject a plea offer, along with some guilty types, while other guilty types accept it. Then, we have $\lambda_I = 1, \lambda_G = \lambda(p_L)$. Moreover, we have

$$\mu' = \frac{\lambda\mu}{\lambda\mu + 1 - \mu} < \mu. \tag{6}$$

The fact that some guilty types reject plea bargain, making $\mu' > 0$, and innocent types also reject plea bargain so that $\mu' < 1$, justifies the costs of holding a trial. A full-separation of types at the plea bargain would have made the trial redundant.

Proposition 4 *In any equilibrium, the plea offer punishment exceeds the following threshold:* $p_L \geq (q - c) \times j$.

3.3 Stages 1 and 2 combined: Reporting of battering and violence decisions

Next, we examine the incentives for women to file false reports of domestic violence. First, we must consider what happens if conditions are such that no true reports are lodged—either because potentially violent men are not being violent, or because battered women are not reporting. In this event, jurors will immediately realize that the defendant is innocent and will automatically acquit without bothering to process costly trial-relevant information. In addition, the jury returns a verdict of *false accusation.*

Assumption 2 *Suppose that in the event the jury returns a verdict of false accusation, the judge can slap the plaintiff with a penalty of* $F > r_f$ *where, recall,* r_f *is the utility to a malicious woman from false accusation.*

The imposition of a penalty for false accusation is a realistic assumption. For instance, in the USA, false reporting is a Class 1 misdemeanor punishable by up to 6 months in jail and $\$2,500$ in fines, while perjury can bring Class 4 felony penalties including 1-3.75 years in jail.[16] Similarly, in India, under Section 248 of the BNS (Bharatiya Nyaya Sanhita),

[16] https://www.salwinlaw.com/penalty-for-false-accusation-of-domestic-violence/

false malicious accusations are punishable with up to two years' imprisonment, fines, or both.[17] This penalty is aimed at deterring false accusation.

One remark about the model formulation at this juncture. Upon receiving a report of domestic violence, the prosecutor first offers a plea bargain and if that is declined by the defendant then she always admits the case to trial; we did not give the prosecutor an action of *throwing out the complaint* (or dismiss a case). In the remainder of the analysis, we continue with this simplifying assumption.

Lemma 2 *(i) As long as some true incidents of domestic violence occur and are reported, some false reports are always lodged.*

(ii) If there are no reports of true incidents of domestic violence (either because $a = \mathsf{NB}$, or $a = \mathsf{B}$ and $\chi_{\mathsf{B}} = \mathsf{N}$), then false reports are never lodged ($\chi_{\mathsf{NB}} = \mathsf{N}$).

Lemma 2 is only a partial summary of the equilibrium that we will report next. We have seen earlier that any equilibrium in which both true and false reports are lodged must be semi-separating. We now make two assumptions that are sufficient (though not necessary) to ensure that (i) battered women always report (i.e., $\chi_{\mathsf{B}} = \mathsf{R}$), and non-malicious women would never choose to do false reporting (see Fig. 2).

Assumption 3 $r - m - w_3 - F > -h$.

Assumption 4 $\mathscr{r} > w_1$ *and* $\mathscr{r}' > w_2$.

Proposition 5 (Violence & malicious reporting) *Suppose Assumptions 1–4 hold. In addition, let condition (4) hold. Moreover, suppose that the plea bargain offered by the prosecutor, p_L, lies in the range*[18]

$$\big[\max\{(q-c)\times j, w_1\},\ p(0.5)\times j\big],$$

$$\textit{and} \qquad p_L < v.$$

Then, a violence equilibrium with both true and malicious, false reporting exists. A more detailed equilibrium characterization is as follows:

(i) In stage 4, jurors pay attention with positive probability $\sigma(\mu')$ solving Eq. (3), and choose a verdict according to majority signals. In case everyone fails to pay attention, the jury acquits the defendant but without asserting any malicious intent on the part of the plaintiff.

[17] The BNS replaced the Indian Penal Code (IPC) in 2024 (https://en.wikipedia.org/wiki/Bharatiya_Nyaya_Sanhita; see also https://www.mha.gov.in/sites/default/files/250883_english_01042024.pdf).

[18] Here onward, we treat w_1 solely as victim compensation.

(ii) *In the continuation game from stage 3 onward, the equilibrium in the plea-bargain stage is semi-separating with the innocent men rejecting the plea offer and opting for trial with probability* $\lambda_I = 1$*, and the guilty men rejecting plea offer with probability* $\lambda_G = \lambda(p_L) = \frac{(1-\mu)p^{-1}(p_L/j)}{\mu\left(1-p^{-1}(p_L/j)\right)} > 0$*. The probability of a correct verdict at trial is* $p = p_L/j$*.*

(iii) *In stage 2, battered women will report (*$\chi_B = R$*). Malicious women who have not been subjected to battering by their non-violent men will do false reporting (*$\chi^m_{NB} = R$*), but non-malicious women will abstain from it (*$\chi^{nm}_{NB} = N$*).*

(iv) *Intrinsically violent men will batter their women (*$a = B$*).*

This result shows that not only violent men would batter their women but simultaneously malicious women also make false accusations. Thus, we have twin problems: A society failing to prevent battering enables malicious women to hide behind the truly victimized women, confounding the problem further. To our knowledge, the earlier literature on domestic violence has not given any importance to the problem of malicious accusations, although law enforcement in different countries clearly recognizes it.

Can the prosecutor set the plea bargain in such a way that the above equilibrium does not arise? We show in the supplementary Appendix (Proposition 9) that if the plea punishment is set sufficiently high, we can support a no-battering, no false-reporting equilibrium. But a natural restriction on (plea) punishment makes such crime deterrence an unrealistic goal.[19]

4 Mandatory arrest: Case of heterogeneous v and h

In the analysis so far, we have assumed that all intrinsically violent men are homogeneous in that if they choose to be violent, they all obtain a fixed payoff v, which, moreover, is common knowledge. We have also assumed that all women subjected to violence face the same harm h, so that battered women receive $-h$. This, too, is common knowledge.

We now make modifications to the above two assumptions. Suppose as before that a fraction $1 - \alpha$ of men is intrinsically violent, but now they differ in v. Each man draws a realization of v from the distribution G (with pdf g) in the interval $[v_{\min}, \bar{v}]$. While the lower limit $v_{\min}$ is strictly positive, all intrinsically non-violent men choose zero violence. Instead of a utility, we can now conceptualize the realization of v as a measure of how violent a particular man is. It is reasonable to assume that the harm a woman suffers from her partner's violence is reflective of the severity of the violence. Here, we assume a very simple form of positive correlation: for a particular couple, $h = -v$. The couple is

[19] Complete deterrence would require $p(0.5)j > v$, or $j > v/p(0.5)$. This might be unrealistic if, for example, penalties are set according to the harm done, say, at $j = v$.

aware of the realization of ν (and therefore h), but others do not know. Thus, prosecutors and jurors only know the distribution of ν. Now, in addition to not knowing whether a particular accused man is innocent (intrinsically non-violent) or guilty, they also do not know the value of ν.

We also modify our assumption on reporting. Assumption 3 ensured that all battered women would report. We drop this assumption.

The timeline of the game remains unchanged (Fig. 1). In Fig. 2, the only change is that $-h$ is replaced by $-\nu$. Moreover, it can be easily checked that the stage 4 and stage 3 continuation games remain unaffected. Stage 4 maps the jurors' posterior belief on guilt given that the defendant has rejected a plea bargain and come for trial, into their choice of the probability of paying costly attention and therefore into the probability of the verdict being correct. For a given μ', the choices of σ and p remain identical to the main model. It is also easy to check that there is neither a separating nor a pooling equilibrium in the stage 3 continuation game. Hence, Propositions 1–4 all hold unchanged, and any equilibrium involving plea bargaining and trial must be semi-separating. Therefore, as before, we can use λ and 1 to denote the probabilities of, respectively, a guilty and an innocent defendant's rejection of a plea bargain; as before, λ will be a function of the terms of the plea offer.

From the game tree, a battered woman's payoff from reporting can be written as:

$$\Gamma(p_L) = r - \lambda(p_L)m + (1-\lambda(p_L))w_1 + \lambda(p_L)p_L\frac{w_2}{j} - \lambda(p_L)(1-\frac{p_L}{j})w_3. \tag{7}$$

In (7), we have used the fact that as any equilibrium with plea bargaining and trial is semi-separating, guilty defendants will be indifferent between acceptance and rejection of a plea bargain, yielding $p_L = p \times j$ or $p = p_L/j$. A battered woman reports if and only if

$$\Gamma(p_L) > -\nu$$

or, equivalently,

$$-\Gamma(p_L) < \nu. \tag{8}$$

We next show that under suitable parameter restrictions, payoff from reporting is increasing in the severity of plea bargain. We maintain these restrictions for the remainder of this paper.

Lemma 3 *Suppose $w_2 - w_1 > m$, and that the elasticity of the probability of plea rejection with respect to the severity of the plea bargain, $\epsilon = \frac{\partial\lambda}{\partial p_L}\frac{p_L}{\lambda} < \frac{(q-c)}{(1-q+c)}$. This is sufficient to ensure that $\Gamma'(p_L) > 0$.*

Note that the upper limit on ϵ imposed by Lemma 3 is not too restrictive given that this limit is strictly above 1 (from Assumption 1).

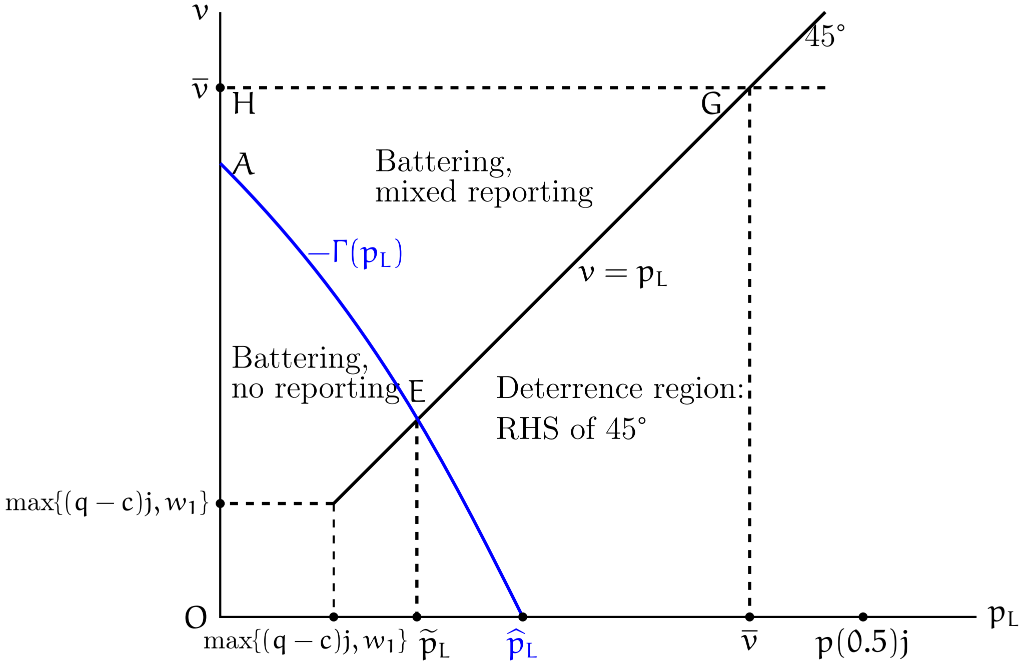


Figure 4: Area $AEGH$ is where both violence and true and false reporting happen. The reason for violence is simple: $p_L < v$ (defendant can always avail plea offer); deterrence is also easy to understand: expected penalty $p \times j = p_L > v$. Payoff from reporting true violence is $\Gamma(p_L)$, and the payoff from not reporting true violence is $-v$, implying there will be truthful reporting if $\Gamma(p_L) > -v$. And so long as there is some truthful reporting, some women will engage in false reporting. But there is a region where battering goes unreported by relatively low-harm victims because plea penalty is not high enough (left of $-\Gamma(p_L)$ curve).

Assumption 5 $\bar{v} > w_1$.

As we will see from the later results, this assumption prevents interesting cases with positive battery and plea bargaining from being ruled out.

Fig. 4 plots two constraints, the reporting constraint (8) and a "deterrence constraint" $p_L > v$, in the (p_L, v) space. As the conditions supporting Lemma 3 hold, the reporting constraint is downward-sloping; since $\Gamma'(p_L) > 0$, we have $-\Gamma'(p_L) < 0$; and the constraint graphs $-\Gamma(p_L)$. Accordingly, if women experience levels of v above the reporting curve, they report; otherwise, they do not. As plea bargains get harsher, women become more willing to report and hence the range of v for which a report of battery will be made expands. We assume that at a particular level of p_L, which we call $\hat{p}_L$, it becomes worthwhile for battered women to report any level of violence. This threshold is defined by $\Gamma(\hat{p}_L) = 0$. We also assume that $\max\{(q-c)j, w_1\} < \hat{p}_L < \min\{\bar{v}, p(0.5) \times j\}$. The lower bound ensures that when the plea bargain is relatively lenient, though in the semi-separating equilibrium range, some women do not report (if not, the problem would

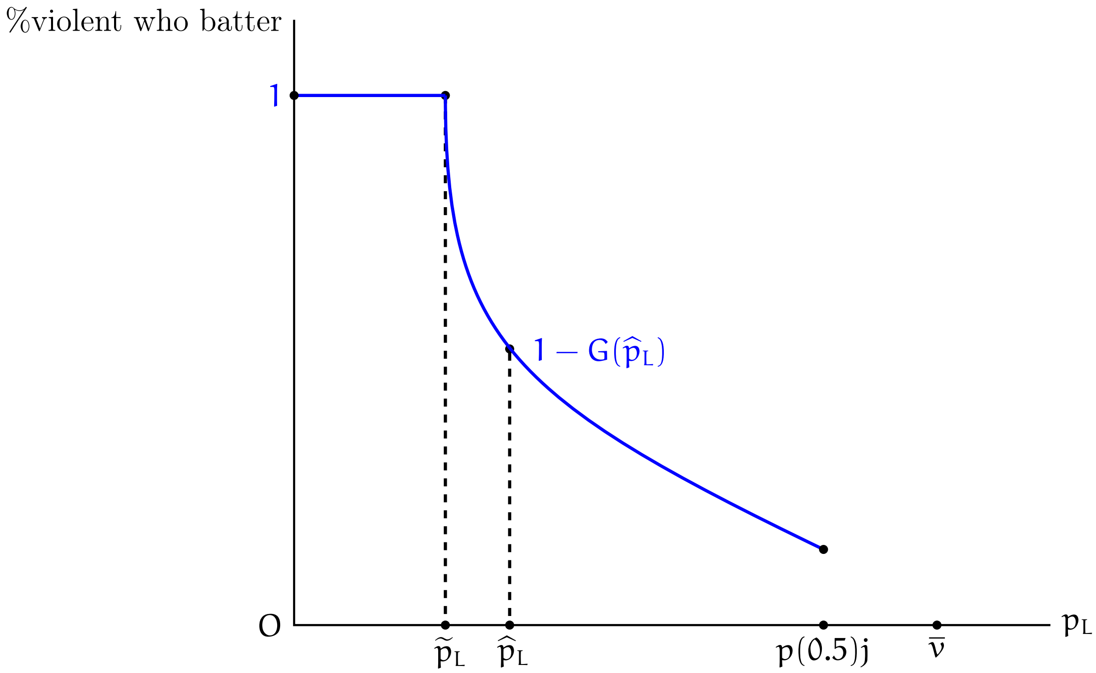


Figure 5: Battery (among violent men) as a function of plea offer. Depending on the level of the plea offer, there is a continuum of battery levels. An initial increase in the severity of plea bargain does not reduce battery among violent men. However, beyond a limit $\widetilde{p}_L$, battering is reduced and continuously decreases in p_L, initially at a higher rate. Thus, any level of plea bargain defines a different level of battery.

become less interesting as the reporting constraint would have no bite). The upper bound ensures that when plea bargains are at the high end of the range, battered women are always ready to report. Finally, in drawing the reporting constraint we have assumed that, at the minimum feasible plea bargain, $\max\{(q - c)j, w_1\}$, some battered women – those experiencing very high violence – are still reporting; that is, $-\Gamma(\max\{(q - c)j, w_1\}) < \bar{v}$.

The deterrence constraint shown in Fig. 4 is a 45°-line in between the permissible limits of the plea bargain. If $v < p_L$, then any man who is sure of being reported (i.e. if v lies above the reporting constraint but below the deterrence constraint) will be deterred, expecting a negative overall payoff from violence, and will not batter his woman. Men above this line will, however, batter their women regardless. In addition, if v lies below both the deterrence and the reporting constraints, potentially violent men will batter: because their violence will anyway not be reported, they obtain a higher payoff, v, from battering than from non-violence. However, note that this range corresponds to low levels of v. We define the plea bargain level at which the two constraints intersect (noting that there is only one intersection as the deterrence constraint is upward-sloping while the reporting constraint is downward-sloping) as $\widetilde{p}_L = \widetilde{v} = -\Gamma(\widetilde{p}_L)$. As is evident from the slopes of the constraints, we must have $\widetilde{p}_L < \widehat{p}_L$. From Fig. 4, we can identify three distinct parameter zones:

1. When $p_L < \widetilde{p}_L$, all violent men will batter. Some are not reported, but even those who are reported all lie above the deterrence constraint, hence they batter regardless.

2. When $\widetilde{p}_L \leq p_L < \widehat{p}_L$, a fraction $1 - G(p_L) + G(-\Gamma(p_L))$ of violent men will batter. This fraction is decreasing in p_L as we can verify from the derivative of the fraction with respect to p_L, which is $-g(p_L) - g(-\Gamma(p_L))\Gamma'(p_L) < 0$. The men who batter are either those who are reported but are still not deterred – those with high ν – or those whose ν is too low to be reported, and who, therefore, feel safe in battering. Men with intermediate levels of ν know that they will be reported and are deterred. Thus, men who commit intermediate levels of violence are the first to drop out (of violence). As plea bargains become harsher, even those with higher ν start getting deterred, while those with lower ν also start getting reported as reporting even low ν starts to become profitable for battered women. Therefore they, too, are deterred.

3. When $p_L \geq \widehat{p}_L$, the fraction of violent men battering falls further to $1 - G(p_L)$, which is also decreasing in p_L, though at a lower rate (the derivative of the fraction is now $-g < 0$). Now, all acts of violence, however small, are reported. Hence, the only men committing battery are those for whom ν exceeds p_L, a range which falls as plea offers get even harsher. The fraction falls to $1 - G(p(0.5)j)$ for $\bar{\nu} > p(0.5) \times j$, as shown in Fig. 5.

Regarding false reports, we maintain the treatment of the previous model without heterogeneity. We retain Assumption 2, and we can verify that Lemma 2 also continues to hold.

Next, we have the following lemma.

Lemma 4 *(i) Within the semi-separating interval where* $\mu' \in (1 - q + c, 0.5]$*, both* λ *and* μ' *are increasing functions of* p_L *and* $\frac{1}{\lambda}\frac{\partial \lambda}{\partial p_L} = \frac{1}{\mu'(1-\mu')}\frac{\partial \mu'}{\partial p_L} - \frac{1}{\mu(1-\mu)}\frac{\partial \mu}{\partial p_L}$.

(ii) When $p_L < \widetilde{p}_L$, $\mu = \frac{(1-\alpha)\left(1-G(-\Gamma(p_L))\right)}{\alpha s+(1-\alpha)\left(1-G(-\Gamma(p_L))\right)}$*, which increases in* p_L*, while it decreases in* p_L *for* $p_L \geq \widetilde{p}_L$*, in which range we have* $\mu = \frac{(1-\alpha)\left(1-G(p_L)\right)}{\alpha s+(1-\alpha)\left(1-G(p_L)\right)}$.

In Proposition 6 below, we present a more nuanced parallel to Proposition 5.

Proposition 6 (High-violence battery, false accusation) *Suppose Assumptions 1, 2, 4 and condition* (4) *hold. Suppose that the plea bargain,* p_L*, offered by the prosecutor is in the range*

$$\begin{aligned} &\left[\max\{(q-c)j, w_1\},\ p(0.5) \times j\right], \\ \textit{and} \qquad &p_L < \bar{\nu}. \end{aligned}$$

Then, an equilibrium with at least some violence and both true and false (malicious) reporting exists, with the extent of violence being a function of the plea bargain. A more detailed characterization follows.

(i) In stage 4, the jurors pay attention with positive probability $\sigma(\mu')$ solving Eq. (3), and choose a verdict according to the majority signals. In case everyone fails to pay attention, the jury acquits the defendant but without asserting any malicious intent on the part of the plaintiff.

(ii) In the continuation game from stage 3 onward, the equilibrium in the plea-bargain stage is semi-separating with the innocent men rejecting the plea offer and opting for trial with probability one, and the guilty men rejecting the plea offer with probability

$$\lambda(p_L) = \frac{1-\mu(p_L)}{\mu(p_L)} \frac{p^{-1}\left(\frac{p_L}{j}\right)}{1-p^{-1}\left(\frac{p_L}{j}\right)},$$

where

$$\mu(p_L) = \frac{(1-\alpha)\big(1-G(\max\{p_L, -\Gamma(p_L)\})\big)}{\alpha s + (1-\alpha)\big(1-G(\max\{p_L, -\Gamma(p_L)\})\big)}.$$

The probability of a correct verdict in the trial is $p = \frac{p_L}{j}$.

(iii) At stage 2, those battered women who experience $\nu > -\Gamma(p_L)$ will report ($\chi_B = R$) while those who experience $\nu < -\Gamma(p_L)$ will not report ($\chi_B = N$). Malicious women who have not been subjected to battering by their non-violent men will do false-reporting ($\chi_{NB}^{m} = R$) while non-malicious women will abstain from it ($\chi_{NB}^{nm} = N$).

(iv) When $p_L < \widetilde{p}_L$, all violent men batter ($a = B$). For $\widetilde{p}_L \leq p_L < \widehat{p}_L$, a fraction $1 - G(p_L) + G(-\Gamma(p_L))$ of violent men batters, the fraction decreasing in p_L. For larger p_L (but smaller than ν), a smaller proportion $1-G(p_L)$ of violent men batters, which also decreases in p_L.

Proposition 6 illustrates that it is harder to deter higher levels of violence under mandatory arrest laws, for a given plea offer. We now turn to an analysis of the situation when arrests require a warrant which must be evidence-based.

5 Arrest with warrant under costly deception

If a woman files a report of DV with the police, the latter incur a fixed cost to investigate. This fixed cost explains why the police do not attempt to investigate under mandatory arrests, where they proceed with all reported cases, without prior investigation. However, now the law requires the police to investigate, because arrests require a warrant, which has to be based on investigation. If the police observe plausible evidence, that is, signal, they enter the case in the police logbook for the prosecution's action.

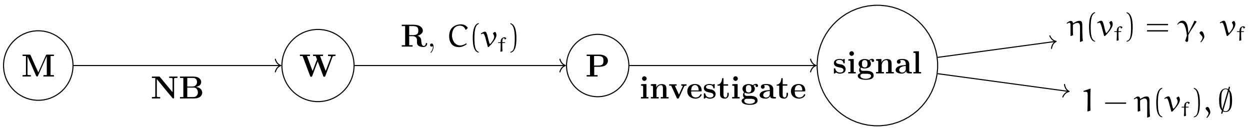


Figure 6: False accusation and police investigation

We assume that they receive such a signal with probability $\eta(\nu)$ when violence of magnitude ν has actually been perpetrated. Moreover, $\eta'(\nu) > 0$ (more violence makes it more likely that a signal is received), and $\eta''(\nu) > 0$, so that the probability of the signal is increasing at an increasing rate in the extent of violence. During the investigation, the police noiselessly infer the actual level of ν (subject to receiving a signal) and report it when they pass the case on to the prosecution.

What happens in case of false reports? Let us denote the magnitude of the violence faked by a malicious woman by ν_f. Then, we assume that if a malicious woman does not fake evidence of violence but nonetheless files a report, the police definitely realize that the case is malicious and no violence has been perpetrated: $\eta(\nu_f = 0, \nu = 0) = 0$. Then the case is thrown out. Thus, if a malicious woman decides to lodge a report, she will choose at the same time to fake evidence of violence of some intensity $\nu_f > 0$. We assume that for all such cases, the police receive a signal that the case is genuine with a constant positive probability γ. Thus, we have $\eta(\nu_f, \nu = 0) = \gamma \ \forall \nu_f > 0$. Further we assume that $\gamma < \eta(\nu_{\min})$, reflecting the investigative ability of the police. When police do obtain a signal in fake evidence cases, they infer the violence at the precise level faked by the woman and enter this level into the case logbook. Manufacturing evidence of violence ν entails a cost $C(\nu)$, where $C'(\nu) > 0$ (it is more costly to fake evidence of a higher level of violence), and $C''(\nu) > 0$ (increasing marginal cost of faking).[20]

If a charge, whether true or false, is forwarded by the police, the prosecutor tailors his plea offer with the knowledge of ν. The plea bargain, if refused, leads to a trial, and the information about the charge (i.e., the level of ν that the man has been accused of) is also received by the jury. The jury's main role here is to sift the true cases from the fake malicious ones, which, as before, requires costly attention.

In terms of Fig. 1, the timeline now has additional stages between stage 2 (where the woman chooses whether to report or not) and stage 3 (where the prosecutor offers a plea bargain). The additional actions taken by a malicious woman are captured in Fig. 6; when reporting, the woman also incurs a cost to fake a particular level of violence; the police investigate, receiving a signal of this faked level of violence, and entering the case in the system, with probability γ, and otherwise throwing out the case. In case of a

[20] We retain Assumption 4 which ensures that non-malicious women never lodge false reports.

true report, the timeline is similar except that the woman does not fake evidence; the police investigate and observe the true ν with probability $\eta(\nu)$, and register the case in the system; otherwise, they do not receive a signal and throw it out. In stage 3, now the prosecutor decides not only the plea offer but also the charge, which accurately reflects the level ν the man has been accused of.

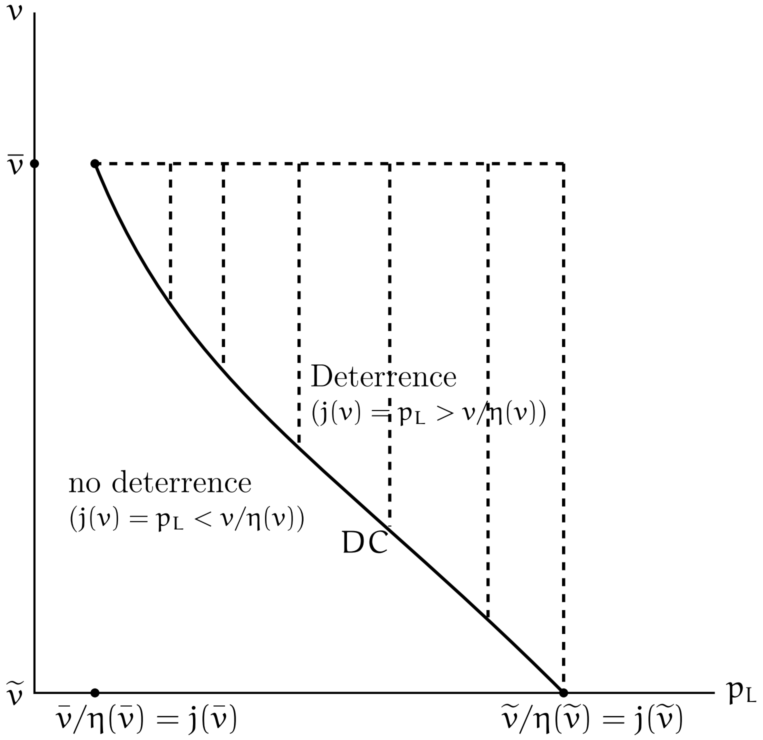

ν
$\bar{\nu}$
$\tilde{\nu}$
Deterrence
$(j(\nu) = p_L > \nu/\eta(\nu))$
no deterrence
$(j(\nu) = p_L < \nu/\eta(\nu))$
DC
p_L
$\bar{\nu}/\eta(\bar{\nu}) = j(\bar{\nu})$
$\tilde{\nu}/\eta(\tilde{\nu}) = j(\tilde{\nu})$

Figure 7: Proposition 7 in picture. DC is the equation binding the deterrence constraint, $j(\nu) = p_L(\nu) \geq \frac{\nu}{\eta(\nu)}$, noting that $\frac{d[\nu/\eta(\nu)]}{d\nu} < 0$ (shown in the proof). Violent men, in committing violence, take into account the probability of being arrested ($\eta(\nu)$). Since no false cases are lodged for $\nu > \tilde{\nu}$, all victims will report. **Summary:** Higher levels of violence are deterred and there is no plea discount ($p_L(\nu) = j(\nu)$).

Proposition 7 (Deterring high violence/false accusation) *(i) False cases involving fake violence ν_f above a threshold $\tilde{\nu} = C^{-1}\big(\gamma \times (r_f + (1-q+c)w_2)\big)$ are never reported, where, recall, the maximum probability of a wrongful conviction upon a false report (with jurors choosing a positive σ^*) is $1 - p_{\min} = 1 - q + c$ as implied by Proposition 2.*

(ii) Genuine incidents of battery involving $\nu > \tilde{\nu}$ can be completely deterred by setting charge-specific jail terms and plea offers $j(\nu) = p_L(\nu) \geq \frac{\tilde{\nu}}{\eta(\tilde{\nu})}$, so there is no plea discount.

This result shows that under the arrest warrant rule, false reporting will be eliminated and any reporting of violence is necessarily truthful. Some reported cases will also be culled out from further consideration due to police not being able to generate any suggestive evidence of battering (a warrant cannot be issued). The rest of the cases will go through the plea bargain and trial process and all defendants receive full punishment commensurate with their crime (no plea discount). Even if some plea offers are refused, jury save on the costs of paying attention as they can convict the defendant with probability one. Thus, judicial costs are minimized.

We may wonder how the above Proposition may be modified if w_2, the compensation received by a woman after obtaining a conviction at trial, were also a function of v, the specific level of violence the man was accused of.[21]

Now, define $\tilde{v}$ as the solution to equation $\tilde{v} = C^{-1}\big(\gamma \times (r_f + (1-q+c)w_2)\big)$, where the argument of $C^{-1}(.)$ is the maximum (gross) benefit from false reporting. The lemma below considers a modified function $w_2(v)$ that is increasing in v, $w_2''(v) \leq 0$ and $w_2(\tilde{v}) = w_2$. In the rest of this section, we consider this newly constructed $w_2(v)$ function.

Lemma 5 *Suppose that* $C'(\tilde{v}) > \gamma(1-q+c)w_2'(\tilde{v})$, *and* $w_2''(v) \leq 0$. *Then, Proposition 7 goes through.*

Lemma 6 *Suppose that both* p_L *and* j *increase in the level of violence that the defendant is charged with. Also, assume that* $\eta(v)g(v)$ *is increasing in* v.[22] *Then,* μ, μ', λ, σ, *and* p *all increase in* v.

In what follows, we make the reasonable assumption that both plea bargains p_L and conviction penalties j are increasing in v that the defendant is charged with. We also maintain the other assumptions supporting Lemma 6.

While Proposition 7 identifies a threshold level of v above which false cases are not lodged, we now ask whether we can uniquely determine an optimal level of v "claimed" or "faked" by malicious women. Let $\pi(v)$ denote the payoff of such a woman from faking evidence of v when an investigation is triggered. Then,

$$\pi(v) = \gamma \times \big[r_f + (1-p(v))w_2(v)\big] - C(v). \tag{9}$$

This leads us to the following result.

Lemma 7 *Suppose the conditions of Lemma 5 hold and that either* $p''(v) \geq 0$, *or* $p''(v) < 0$ *but* $\gamma\big|p''(v)\big| \times w_2(v) < C''(v)\ \forall v < \tilde{v}$. *Then, these restrictions are sufficient to ensure a unique maximizer,* $v^* \in [v_{\min}, \tilde{v}]$, *for* (9).

[21] For example, in California DV cases, the amount of court-mandated victim restitution varies with the seriousness of the charge, being different for felonies and misdemeanors, see https://www.sddvattorney.com/blog/101-what-is-victim-restitution-in-domestic-violence-cases.

[22] For instance, this restriction is obeyed if $G(v)$ is uniform. More generally, the restriction is equivalent to assuming that if $g'(v) < 0$, the absolute value of the elasticity of $g(v)$ with respect to v be smaller than the elasticity of $\eta(v)$ with respect to v, where, recall, this latter elasticity exceeds 1 due to the convexity of $\eta(v)$. If $g'(v)$ is constant or positive, the restriction is automatically satisfied without any assumptions on elasticities.

Lemma 7 has interesting implications: if ν^* is the only level of violence faked by malicious women, then reports of all other levels of violence must involve true battery. The logic of part (ii) of Proposition 7 can then be extended to all these levels of violence, as follows. For the same reason as in part (ii) of Proposition 7, these violence levels can be deterred simply by setting both $p_L(\nu)$ and $j(\nu)$ equal to $\nu/\eta(\nu)$, and convicting with probability one, because it can be inferred that there is no false reporting. (More specifically, since $\nu/\eta(\nu)$ decreases in ν (see proof of Proposition 7), deterrence would be more feasible for relatively high levels of ν (that is, less strict plea bargains and penalties are sufficient to deter relatively high ν). Also, for the same reason as in the Proposition, these cases of battery would always be reported if the man deviates and batters. Hence, all these levels of violence can be deterred, but it is more practical to deter higher levels of violence. No plea discount is offered for any level of violence that is not faked by malicious women.[23])

What happens when reports of violence level ν^* enter the system? We first verify whether it is indeed optimal for battered women experiencing ν^* to report. The expected payoff from a report would be

$$\Gamma(\nu^*) = \eta(\nu^*)\left[r - \lambda(\nu^*)m + (1-\lambda(\nu^*))w_1 + \lambda(\nu^*)p(\nu^*)w_2 - \lambda(\nu^*)(1-p(\nu^*))w_3\right].$$

By assuming the same restrictions as in Lemma 3, we can show that $\Gamma(\nu^*)$ is increasing in p_L and hence in ν^*. Battered women will report ν^* if $\Gamma(\nu^*) > -\nu^*$, i.e., if $-\Gamma(\nu^*) < \nu^*$. Since the LHS of this inequality is decreasing in ν^* while its RHS is increasing in it, reporting is more likely to happen if ν^* is high. Specifically, let some $\underline{\nu}$ solve $-\Gamma(\underline{\nu}) = \underline{\nu}$. Then, ν^* is reported if and only if $\nu^* > \underline{\nu}$.

Proposition 8 (Plea discounts) *(i) If $\nu^* > \underline{\nu}$, then both true and fake reports of ν^* enter the system unless the plea offer is set at $p_L(\nu^*) \geq \frac{\nu^*}{\eta(\nu^*)}$, and $j(\nu^*) \geq \frac{\nu^*}{\eta(\nu^*)p(\nu^*)}$. If $p_L(\nu^*)$ and $j(\nu^*)$ can be set at these high levels, neither true nor false reports enter the system.*

(ii) Plea discounts are now necessary to sustain violence in equilibrium for $\nu = \nu^*$: $p_L(\nu^*) < j(\nu^*)$.

(iii) Deterrence is easier to achieve if ν^* is high.

(iv) If $\nu^* < \underline{\nu}$, deterrence is not achieved, but false reports do not enter the system either.

[23] We may allow malicious women to have different levels of malice, which depend differently on ν_f (that is r_f can differ, and depend differently on ν_f, for different malicious women). These would then enter the expression for the first derivative of fake payoffs w.r.t. reported violence ν_f (Eq. (18)) differently and result in different optimizers for different malicious women. Thus, more than one level of violence could be faked.

Table 1: Comparison of two trial procedures

	Mandatory arrests	Evidence (warrant) based arrests
1.	All false cases enter as long as some true cases are reported	Only $\eta(0)$ of false cases enter
2.	Prob. of guilt in the arrestee pool: $\mu = \frac{(1-\alpha)\left[1-G\left(\max\{p_L,-\Gamma(p_L)\}\right)\right]}{(1-\alpha)\left[1-G\left(\max\{p_L,-\Gamma(p_L)\}\right)\right]+\alpha s}$ $\Big[\mu$ is low, lower than $\frac{1-\alpha}{1-\alpha+\alpha s}$ → known fixed harm $\mu\Big]$	$\mu = 1$ if $\nu \neq \nu^*$; for $\nu = \nu^*$, $\mu(\nu) = \frac{\eta(\nu)g(\nu)(1-\alpha)}{\eta(\nu)g(\nu)(1-\alpha)+\gamma\alpha s}$
3.	Plea discount needed	Plea discount not needed if $\nu > \widetilde{\nu}$, or more generally if $\nu \neq \nu^*$
4.	Uniform penalties	Penalties tailored to ν
5.	Deterrence of high ν is hardest	Deterrence of high ν is easiest
6.	No way of filtering out any false charges	High-violence charges are never faked
7.	Jurors' information gathering costs substantial	Evidence gathering by police mostly substitutes out costly jury deliberation

■ **Summary.** In results from Proposition 7 onwards, we have shown that if the prosecutor had complete freedom over plea bargains, it is possible to deter all levels of ν. Proposition 7 and the discussion following Lemma 7 showed that all levels of violence ν except for ν^* would involve only true cases and could be deterred by high enough (plea) penalties and automatic conviction subject to reaching the trial stage. Proposition 8 then showed that even ν^* could be deterred if penalties were high enough (subject to plea discounts being offered). However, in reality prosecutors may not be able to set very high plea offers, and there may be upperbound restriction on $j(\nu)$. Thus, the results also highlight that for any given plea offer, the highest levels of ν are easier to deter than the lower levels (even if the plea offer is not high enough to achieve complete deterrence of all levels of ν). Moreover, we can avoid giving plea discounts or burdening the jury with the cost of reviewing evidence whenever a report comes in of a violence level that would not be faked.

■ **Comparison between mandatory arrest and arrest with warrant.** Under warrantless (or mandatory) arrest, making false reports is costless and the legal system has had to deal with an abundance of cases—the entire range of ν. The police did little to no investigation to learn about the extent of violence, nor did it therefore bring charges proportionate to the crime. The plea offer is uniform, as is the penalty on conviction. The power to discipline deviant behavior was limited.

In comparison, under the prior warrant requirement, the police would investigate and learn more precisely the extent of battering and bring in the charge according to the level of the crime. Also, the false accusers will have to incur their own costs to fake the accusation. The burden of faking thus disciplines malicious accusers. Warrantless arrest admits too many accusations from all malicious women, clogging up the trial system. This makes disbursement of justice very imperfect, inducing more crimes.

In mandatory warrantless arrest, our model permits only a uniform plea bargain (and penalty) for all cases of violence, as the level of ν is not known from prior reports; jurors are simply instructed to find out whether violence has been committed, or not. As we saw, this implies that men committing high ν are much less likely to be deterred (by any given plea offer) than other men (and indeed this is clear from the shape of the deterrence constraint). (Complete deterrence would have required $p(0.5)j > \bar{\nu}$, which would have implied that the common punishment for any level of ν would have to be higher than a multiple of the highest possible level of violence (given $p(0.5) < 1$).) See Table 1.

Now, suppose we allow the prosecutor to observe a noisy signal of the level of violence (this does not play any role in arrest with a warrant, because the prosecutor already knows the exact level of violence that the defendant has been accused of). For example, if the true level of violence is ν, the prosecutor might (with equal likelihood) observe any realization in the interval $[\nu - \delta, \nu + \delta]$. However, if the true ν is in the upper range of possible violence levels $\bar{\nu} - \delta$ or above, the prosecutor observes any realization in the interval $[\nu - \delta, \bar{\nu}]$ where $\nu - \delta \geq \bar{\nu} - 2\delta$. Similarly, if the case is actually false, the true level of violence is zero, but the prosecutor observes any realization in the interval $[0, \delta]$. The prosecutor sets a plea bargain according to the realization she has observed and also communicates his realization to the judge, who can tailor j according to this realization. Then, although the plea bargain, and the punishments, are no longer uniform, it is still the case that the most severe violence levels are under-deterred relative to others; while all false cases continue to enter the system. Intuitively, for the most severe cases, prosecutors will receive a signal less than the true violence level; therefore, (plea) punishments will be pegged at a lower level. In contrast, if the true violence level were smaller, the defendant might expect that on average he will get the punishments pegged to this true violence level. False cases all enter the system and there is an incentive to file false cases, both due to malice utility and due to possible extra benefits in cases of wrong convictions. While in the interest of simplicity, we have not explicitly worked out this extension, it seems that allowing the prosecutor to observe a noisy signal, while allowing for non-uniform penalties even in the mandatory case, would retain the features that the most severe crimes are relatively under-deterred, and that false cases are a much more prevalent problem, as compared to the regime where arrests are based on warrants or compulsory evidence-gathering by police.

■ **Comparison of judicial costs.** An interesting point relates to the fact that judicial costs are much lower with warrant-based arrests as compared to mandatory, warrantless ones. The primary reason for this is that faking evidence is not worthwhile for many levels of violence, so that reports of such violence levels can be treated by jurors as true, and they may convict without costly information processing. This is something which is not possible with mandatory arrests, as all false reports go through, and hence the jury needs to exert costly effort to distinguish genuine cases from false ones. However, one could argue that costs are being shifted from the judiciary to the police (investigators) in the warrant-based regime, as the police have to undertake costly investigations when reports are lodged. Nonetheless, observe that the police only have to investigate when reports are lodged; if the threat of investigation is credible, false reports are actually deterred for large ranges of violence, while real battery is also deterred for many ranges. Thus, the actual costs incurred by the investigators may not be as large as may appear at first glance.

6 Conclusion

As a central focus, we examined how mandatory (or warrantless) arrests can undermine evidence collection critical to deciding whether a case of domestic violence deserves a serious hearing and whether it may work against the deterrence objective. Existing empirical research does not offer any definitive conclusion: homicides (an extreme form of battering) can increase (Iyengar, 2009) or not (Chin and Cunningham, 2019). We contribute to this research by studying the hypothesis that some alleged victims might be providing false evidence of injuries from malice. It is the possibility of fake allegations that makes the jury's role in distinguishing guilty from innocent defendants meaningful, yet this aspect was not previously modeled in the DV literature.

Our analysis can conceivably be extended to tackle other contexts where genuine wrong-doing coexists with false accusations of wrong-doing. An instance that comes to mind relates to the "Me Too" movement. While exposing many genuine incidents of sexual misconduct, the movement also made it easy to make maliciously motivated accusations. We reserve such themes for future research.

Appendix

Proof of Proposition 1. (i) Plot the LHS and RHS of Eq. (3) on a graph with σ on the x-axis. At $\sigma = 1$, LHS lies below RHS (by condition (4); the reason is that even if everyone is paying attention $(n-1)$ jurors, other than the candidate juror whose strategy is being considered, may still be tied if half receive wrong signals and half receive correct signals). At $\sigma = 0$, $\mu' > 1 - q + c$ implies that LHS of (3) is above RHS. Given that the LHS expression is continuous in σ, it follows by the intermediate value theorem that there exists a solution σ^* to Eq. (3) for any $\mu' \in (1 - q + c, 0.5]$. Also, because LHS starts above RHS and ends below it, there will be an odd number of intersections.

(ii) First consider the case of a unique solution as shown in Fig. 8. The blue curve in Fig. 8 shows that LHS will move up as μ' increases and thus will increase the equilibrium σ^*.

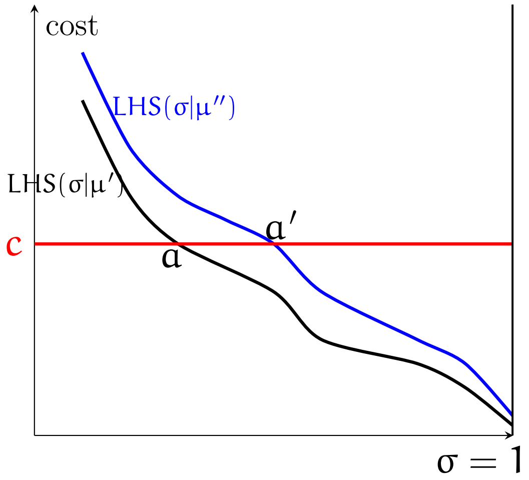


Figure 8: Consider Eq. (3). For $\mu'' > \mu'$, $\mathsf{LHS}(\sigma|\mu'') > \mathsf{LHS}(\sigma|\mu')$. Thus, the unique solution $(\sigma^*(\mu')$ at which $\mathsf{LHS}(\sigma|\mu') = c)$ will increase as μ' increases.

Consider next an odd number of solutions, for example, solutions a, b and d in Fig. 9. Note that b is unstable. Suppose σ falls slightly below its equilibrium value at b. The LHS then falls below the RHS, so there is an incentive to further lower σ, as the cost c will outweigh the expected benefit of paying attention. Restricting to stable equilibria, LHS must cut RHS from above. And from the upward shift of the LHS curve resulting from an increase in μ', we see that any local perturbation of the stable equilibria (a and d) involves a rise in the new equilibrium σ as μ' increases. Finally, the monotonicity of σ^* follows applying juror's risk-dominant preferences.[24]

[24] In the Supplementary Appendix, we include some simulations of σ^* and p for different values of

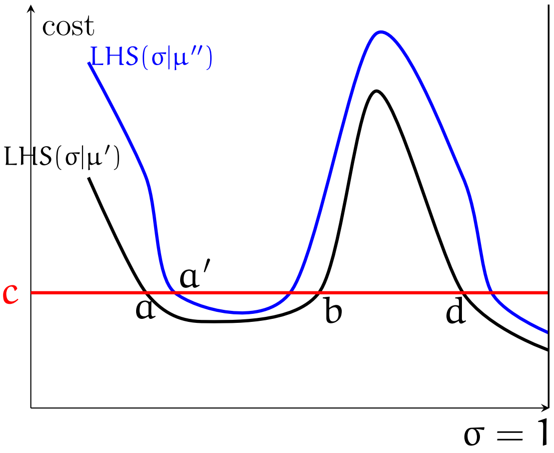


Figure 9: Consider Eq. (3). For $\mu'' > \mu'$, $\text{LHS}(\sigma|\mu'') > \text{LHS}(\sigma|\mu')$. Thus, the risk-dominant solution (minimum $\sigma^*(\mu')$ at which $\text{LHS}(\sigma|\mu') = c$) will increase as μ' increases.

(iii) Denote the utility of a typical juror at the symmetric equilibrium (σ^*) (symbol $(.)$ is for vector) as

$$U = p((\sigma^*)) - \sigma^* c.$$

We know from part (ii) that symmetric equilibrium (σ^*) is increasing in μ'. So consider $d\mu' > 0$ leading to $(d\sigma^*) > (0)$. Denote the new symmetric equilibrium by $(\sigma^{**}) = (\sigma^* + d\sigma^*) > (\sigma^*)$.

Let all but juror i choose σ^{**}. Juror i's f.o.c. is

$$\frac{\partial U}{\partial \sigma_i} = \frac{\partial p(\sigma^{**}, (\sigma^{**})_{-i})}{\partial \sigma_i} - c = 0 \Rightarrow \frac{\partial p(\sigma^{**}, (\sigma^{**})_{-i})}{\partial \sigma_i} = c > 0.$$

We can now write $p((\sigma^*(\mu')))$ and then taking the derivative obtain:

$$\frac{dp}{d\mu'} = \sum_{j=1}^{n} \frac{\partial p}{\partial \sigma_j} \frac{d\sigma_j}{d\mu'} = c \sum_{j=1}^{n} \frac{d\sigma_j}{d\mu'} = cn \frac{d\sigma^*}{d\mu'} > 0. \qquad \textbf{Q.E.D.}$$

Proof of Proposition 2 . The value of μ' at which jurors just stop paying attention, setting $\sigma = 0$ in Eq. (3), is $\mu' = 1 - q + c$ (for $\mu' \leq 0.5$). Now, jurors choose a verdict by consensus based on their updated belief, leading to acquittal, which is only correct with probability $1 - \mu'$, the probability that the defendant is innocent. At the point $\mu' = 1 - q + c$, we therefore have $p = 1 - \mu' = q - c$. At lower μ', with jurors not paying attention, $p = 1 - \mu'$ and thus p will increase as μ' decreases. At higher μ', jurors pay attention, implying that p necessarily increases beyond the level at which jurors only rely on updated belief, since the signal accuracy q always exceeds $1 - \mu'$. By part (iii) of

μ' and jury size, n. For the parameters we examine, σ^* is unique, given μ' and other parameters. Nonetheless, we have explored the possibility of multiple equilibria in this proof because we also found that the derivative of the LHS of Eq. (3) with respect to σ^* is not necessarily always negative, thus raising the possibility of multiple solutions to Eq. (3).

Proposition 1, $p(\mu')$ is strictly increasing. Hence, $p_{min} = q - c$ at the unique minimizer $\mu' = 1 - q + c$, and the V-shape of $p(\mu')$ immediately follows. See Fig. 3. **Q.E.D.**

Proof of Lemma 1. If guilty defendants reject a plea offer and go to trial, they are convicted whenever jurors are correct, while if innocent defendants reject an offer and opt for trial, they are only convicted if jurors are wrong. However, from Proposition 2 we know that jurors are more likely to be correct than wrong (as $p_{min} = q - c > 0.5$), and so the guilty has more to lose by rejecting a plea offer. Hence, the result follows. **Q.E.D.**

Proof of Proposition 3. We first claim that there is no pooling equilibrium in which all defendants, whether they are guilty or innocent, accept a common plea offer p_L. Suppose that such an equilibrium exists. First, by Lemma 1, guilty defendants are less likely to reject a plea offer than innocent ones. Therefore, jurors reason that if a pooling equilibrium exists, it indicates that even innocent defendants do not prefer trial to accepting the plea offer. This then implies that the guilty must strictly prefer to accept the plea offer. Consequently, if jurors observe someone deviating from the pooling equilibrium and show up in the trial, they appeal to the Cho-Kreps criterion and place zero weight on this defendant being guilty. Therefore, they think that $\mu' = 0$, optimally choose $\sigma = 0$ and acquit such a defendant. But then innocent defendants should deviate from the pooling equilibrium as a sure acquittal is more attractive than any plea offer carrying a non-zero penalty. Thus, the posited pooling equilibrium also breaks down.

Next, we show that there is no fully separating equilibrium where all violent (guilty) defendants accept a plea offer p_L, while all innocent defendants (the falsely charged non-violent men) reject it. Suppose, to the contrary, that such a separating equilibrium exists. Then, all jurors think that $\mu' = 0$, will optimally choose $\sigma = 0$ and acquit anyone coming to trial. But then any guilty (violent) man should reject the plea offer mimicking innocent defendants and get acquittal, a contradiction.

Finally, we cannot have a separating equilibrium in which a plea offer is accepted by all innocent types and rejected by all guilty types, since by Lemma 1, guilty types are always more willing to accept a plea offer than innocent types. **Q.E.D.**

Proof of Proposition 4. Suppose that there is an equilibrium plea offer of $p_L < (q - c) \times j$. By Proposition 2, $(q - c)$ is the minimum p, the probability of a correct verdict, which can be induced in equilibrium. So, we have $p_L < p \times j$, that is, the plea-offer punishment is less than the expected cost of going to trial for a violent man who has been reported; he will always accept the plea deal. This implies that only innocent men end up in the trial. But by Proposition 3, a fully separating equilibrium cannot exist. This is a contradiction. **Q.E.D.**

Proof of Lemma 2. (i) If some true incidents are reported in the putative equilibrium, the jurors do not automatically acquit. In this case, if a malicious woman makes a false accusation, she obtains either r_f (if the man is acquitted—clearly no assertion can be made about the woman's malicious intent for the equilibrium under consideration), $r_f + w_2$ (if the man is mistakenly convicted, in which case the woman is compensated), or $r_f + w_1$ (in the out-of-equilibrium event that an innocent man accepts a plea offer). If she does not lodge a false report, she receives zero. Since all her possible payoffs from lodging a false report are positive, she would engage in malicious reporting.

(ii) If no true incidents are reported in the putative equilibrium, the jury immediately acquits the man and returns a verdict of malicious intent so that the woman will be asked to pay a penalty,[25] so her payoff is $r_f - F < 0$ (by Assumption 2). Thus, she prefers not to lodge a false report. **Q.E.D.**

Proof of Proposition 5. [(i)-(ii)]. For the time being, assume $\mu' \in (1 - q + c, 0.5]$. Then, Assumption 1 together with the condition (4) establishes the claim in part (i), by Proposition 1. In the proof of part (ii) below, we will verify the assumed range of μ'.

By Proposition 3, the equilibrium in the plea-bargain stage cannot be separating or pooling, and hence must be semi-separating. Thus, guilty defendants (violent men) must be indifferent between accepting a plea bargain and going to trial. Accordingly, their (residual) continuation payoff from acceptance, $-p_L$, must equal their expected payoff from trial, which is $-(p \times j)$, implying the probability of a correct verdict

$$p = p_L/j. \tag{10}$$

Recall from Proposition 1 (and here we continue to assume $\mu' \in (1 - q + c, 0.5]$, so we can invoke Proposition 1) that p is an increasing function of μ': $p = p(\mu')$, $p'(.) > 0$. By inverting,

$$\mu' = p^{-1}(p(\mu')) = p^{-1}(\frac{p_L}{j}), \tag{11}$$

where the second equality follows from (10). Using (11) in (6) and manipulating, we obtain

$$\lambda_G = \lambda = \frac{(1-\mu)p^{-1}(\frac{p_L}{j})}{\mu\big(1 - p^{-1}(\frac{p_L}{j})\big)}. \tag{12}$$

Since $p_L \leq p(0.5) \times j$, we have

$$\frac{p_L}{j} \leq p(0.5) \quad \text{or,} \quad p^{-1}(\frac{p_L}{j}) \leq 0.5,$$

[25] Note that if true cases are also being lodged, jury will never slap a penalty on a woman whose partner has gone to trial and been acquitted. This is because this could also happen due to judicial inattention. Thus, it does not necessarily imply that the report was false.

because $p(\mu')$ is an increasing function (Proposition 1).

Now go back to the first equality in Eq. (11) and in the argument insert the solution $p(\mu')$ from Eq. (10) to obtain:

$$\mu' = p^{-1}(\frac{p_L}{j}) \leq 0.5, \tag{13}$$

thus verifying the previously assumed upper bound of μ' in $\mu' \in (1 - q + c, 0.5]$.

Examining the incentives of innocent men to reject the plea offer, we find that doing so gives them an expected payoff of $-m_f$–$(1 - p) \times j$ (the second term capturing the expected punishment from a wrongful conviction), while if they accept it, they obtain $-m_f - p_L$. Now, we have $p_L \geq (q - c) \times j > (1 - q + c) \times j \geq (1 - p) \times j$. The first (weak) inequality is the assumed lower bound on p_L; see the proposition statement. The second inequality follows from Assumption 1. The third (weak) inequality follows from Proposition 2 (specifically, $p_{min} = q - c$). Thus, we have $-m_f - p_L < -m_f$–$(1 - p) \times j$, so innocent men reject the plea bargain with probability one.

We know that jurors will pay attention by choosing positive σ as long as $\mu' \geq 1-q+c$. From (11), the lowest possible value of μ' is where $p_L = (q - c) \times j$, and $p = q - c$, which occurs at the point where $\mu' = 1 - q + c$. This verifies the assumed lower bound of μ'.

Finally, the jury returning a verdict according to the majority signal is clearly optimal. When none of the jurors observe a signal, their acquittal decision follows from the fact that $\mu' \leq 0.5$.

(iii) Assumption 3 guarantees that battered women always report. To see this, note that a battered woman's payoff if she does not report is $-h$. If she does report, her expected payoff is $\lambda[p(r - m + w_2) + (1 - p)(r - m - w_3)] + (1 - \lambda)(r + w_1)$; note that when the verdict is inaccurate, the probability of which is $1 - p$, the jury can only acquit the defendant without asserting that the plaintiff is malicious and thus the lowest branch following the trial node in Fig. 2 does not come into play for a battered woman along the equilibrium path (and hence the payoff $r - m - w_3 - F$ is not applicable). Assumption 3 ensures that even when the battered woman fails to secure a conviction, she should still prefer her payoff to one from not reporting: $r - m - w_3 > -h$. Now, the payoffs from reporting are strictly higher if the man accepts a plea bargain because of the compensation that the woman receives or if she obtains a conviction. Thus, $\chi_B = R$. But then, from part (i) of Lemma 2 it follows that the woman *may* lodge a false report. Because the lemma does not decisively pin down which type of women will do false reporting, we need to analyze the decision for each type separately, the malicious and non-malicious types. By submitting a false report, the malicious type receives $(1-p)(r_f+w_2)+pr_f > 0$ (recall that an innocent man will reject plea bargain with probability 1, so look at the plaintiff's payoffs following two possible jury decisions, convict and acquittal, in the lower branch of Fig. 2; the jury never chooses acquittal with a verdict of malicious intent on the part of the plaintiff) which clearly dominates the payoff of 0 from not reporting. So, the malicious woman will report falsely when not battered. On the other hand, by reporting falsely the non-malicious type receives $(1-p)(-\mathscr{r}' + w_2) + p \times (-\mathscr{r}'') < 0$, because by Assumption 4 $\mathscr{r}' > w_2$. But by not reporting she will receive 0. Hence, non-malicious type will choose not to lodge a false report.

(iv) Finally, given that $p_L < v$, violent men will choose $a = B$ and not deviate. Although their violence will be reported, their expected payoff from battering and subsequent punishment is $v - p_L$ (this is the case even for men who reject the plea offer and go to trial, as the expected penalty at trial is equal to the plea bargain by Eq. (10)), and this is positive. Thus, it exceeds their payoff from not battering which is at most zero. **Q.E.D.**

Proof of Lemma 3. Differentiating (7) with respect to p_L, we see that $\Gamma'(p_L) > 0$ if and only if

$$\frac{\partial \lambda}{\partial p_L}(m + w_1 + w_3) < \left\{ \lambda + p_L \frac{\partial \lambda}{\partial p_L} \right\} \left(\frac{w_2 + w_3}{j} \right)$$

or,

$$\frac{\lambda \epsilon}{p_L}(m + w_1 + w_3) < \lambda(1+\epsilon) \left(\frac{w_2 + w_3}{j} \right).$$

By canceling common terms, simplifying, and substituting $p_L/j = p$, this is equivalent to

$$\frac{m + w_1 + w_3}{w_2 + w_3} < p \frac{(1+\epsilon)}{\epsilon}.$$

Now, the LHS of the above inequality is always smaller than 1, as $m + w_1 < w_2$. The RHS is always larger than $(q-c)\frac{(1+\epsilon)}{\epsilon}$ as, by Proposition 2, $q - c$ is the smallest feasible value of p. Now, this RHS is therefore also always larger than 1 given that we have $\epsilon < \frac{q-c}{1-q+c}$. The result follows. **Q.E.D.**

Proof of Lemma 4. (i) Suppose that p_L increases and suppose that, to the contrary, μ' decreases (stays constant). Then, as we are in the interval $\mu' \in (1-q+c, 0.5]$, we can invoke Proposition 1. Therefore, p decreases (stays constant). Since p_L increases while $p \times j$ decreases (stays constant), this induces guilty defendants to all reject the plea bargain and go to trial, which contradicts the semi-separating nature of the equilibrium. Thus, μ' increases in p_L. That λ increases in p_L follows from the fact that a more severe plea bargain has a larger chance of being rejected. Next, note that we have

$$\lambda = \frac{\mu'(1-\mu)}{\mu(1-\mu')}.$$

Differentiating both sides with respect to p_L, we have

$$\frac{\partial\lambda}{\partial p_L} = \frac{1-\mu}{\mu}\frac{\partial\mu'/\partial p_L}{(1-\mu')^2} - \frac{\mu'}{1-\mu'}\frac{\partial\mu/\partial p_L}{\mu'^2}.$$

Using the definition of λ and rearranging, we get

$$\frac{1}{\lambda}\frac{\partial\lambda}{\partial p_L} = \frac{1}{\mu'(1-\mu')}\frac{\partial\mu'}{\partial p_L} - \frac{1}{\mu(1-\mu)}\frac{\partial\mu}{\partial p_L}. \tag{14}$$

(ii) When $p_L < \widetilde{p}_L$, $\mu = \frac{(1-\alpha)\left(1-G(-\Gamma(p_L))\right)}{\alpha s+(1-\alpha)\left(1-G(-\Gamma(p_L))\right)}$. Reports of domestic violence are made either by malicious women of non-violent men (αs such reports are made) or by women of violent men who experience sufficient violence to satisfy the reporting constraint, those who experience violence greater than $-\Gamma(p_L)$. Thus, the proportion of guilty defendants in the pool is given by the expression above. Differentiating with respect to p_L, we obtain

$$\frac{\partial\mu}{\partial p_L} = \frac{-(1-\alpha)\alpha s g \Gamma'(p_L)}{\left[\alpha s + (1-\alpha)\left(1-G(-\Gamma(p_L))\right)\right]^2} > 0. \tag{15}$$

When $p_L \geq \widetilde{p}_L$, $\mu = \frac{(1-\alpha)\left(1-G(p_L)\right)}{\alpha s+(1-\alpha)\left(1-G(p_L)\right)}$. Those intrinsically violent men who are reported are the ones for whom ν lies above both the deterrence and the reporting constraints, that is, a fraction $1-G(p_L)$ of intrinsically violent men. The remaining reports are made by the malicious wives of non-violent men. Differentiating with respect to p_L, we obtain

$$\frac{\partial\mu}{\partial p_L} = \frac{-(1-\alpha)\alpha s g}{\left[\alpha s + (1-\alpha)\left(1-G(p_L)\right)\right]^2} < 0. \tag{16}$$

From (14), (15), and (16) we see that λ increases more sharply in p_L when $p_L \geq \widetilde{p}_L$, reflecting both the rise in μ' and the fall in μ. Its increase when $p_L < \widetilde{p}_L$ is more modest reflecting the fact that while μ' must still rise with p_L, so does μ. **Q.E.D.**

Proof of Proposition 6. Parts (i) and (ii) of this proof are identical to the proof of the corresponding parts of Proposition 5. The definition of μ follows from Lemma 4 part (ii). For part (iii) of the proof, the behavior of non-battered women (both malicious and non-malicious) exactly mimics that of the behavior of such women in part (iii) of the proof of Proposition 5. For battered women, the reporting constraint by definition demarcates women experiencing ν greater than $-\Gamma(p_L)$, who prefer reporting to not reporting, and those below it, who prefer not to report. (Note that because of our earlier assumptions, every battered woman reports when p_L is high enough, that is above $\widehat{p}_L$). For part (iv), when $p_L < \widetilde{p}_L$, violence is either not reported – in which case it is profitable to batter as it yields a payoff of ν – or, even if it is reported, the expected payoff $\nu - p_L$ is always positive as any ν which lies above the reporting constraint also lies above the deterrence constraint in this range. Thus, no violent men are deterred. When $\widetilde{p}_L \leq p_L < \widehat{p}_L$, men with intermediate levels of ν are deterred. These levels of ν are above the reporting constraint but below the deterrence constraint; aware that they would be reported if they batter, these men prefer not battering to the negative payoff $\nu - p_L$. The range of men deterred goes up with p_L. Above $\widehat{p}_L$, all acts of violence are reported, so all men with $\nu < p_L$ are deterred, and this fraction goes up with p_L. However, as p_L is less than $\bar{\nu}$, some violence happens for high ν men, who then go through the justice process. **Q.E.D.**

Proof of Proposition 7. (i) If an investigation is launched into a case where violence has actually not occurred, the maximum gain that the woman expects by manufacturing false evidence of violence is

$$\gamma(r_f + (1-q+c)w_2) \geq \gamma(r_f + (1-p)w_2).$$

This follows from the fact that the minimum value of p is $q - c$ (Proposition 2), and hence $1 - p$ is less than or equal to $1 - q + c$. The woman expects that, if her fake case enters the system (with probability γ) that she will receive her payoff from malice (which she does not if her case is thrown out), and if the man is incorrectly convicted she expects to receive additional compensation w_2. (The man being innocent, never accepts a plea bargain if he is charged, from previous results.) Against this possible maximum gain, the woman incurs a cost $C(\nu)$ to manufacture fake evidence of a level of violence ν. Therefore, she will never find it worthwhile to produce evidence of too high a level of violence such that the cost of faking exceeds the maximum possible gain. Hence, false cases involving

$$\begin{aligned} C(\nu) &> \gamma\big[r_f + (1-q+c)w_2\big] \\ \text{or,}\quad \nu &> \widetilde{\nu} = C^{-1}\big(\gamma(r_f + (1-q+c)w_2)\big), \end{aligned}$$

are never lodged.

(ii) From part (i), it is known that false cases are never lodged in this range of ν. Thus, if an investigation uncovers evidence of $\nu > \widetilde{\nu}$, and is passed on to the prosecutors and the jury, guilt is definite. Hence, if the case reaches trial, the jury will convict with probability one without even bothering to incur c, costly information processing. Since the probability of conviction is 1, the prosecutor will also offer a plea bargain exactly equal to the penalty on conviction. Notice that a battered woman subjected to $\nu > \widetilde{\nu}$ will always report as her expected payoff from doing so is either $\eta(\nu)[r + w_1] > 0 > -\nu$ (if the plea bargain is accepted) or $\eta(\nu)[r + w_2 - m]$ (which is greater than $\eta(\nu)[r + w_1]$, as $w_2 - w_1 > m$, as assumed in Lemma 3) if the case goes to trial, given that the conviction is automatic.

Check that $\frac{v}{\eta(v)}$ is decreasing in v: $\frac{\partial[v/\eta(v)]}{\partial v} = \frac{\eta(v)-v\eta'(v)}{\eta(v)^2} < 0$ where the negative sign follows from the convexity of $\eta(v)$. Hence, if $j(v)$ for all $v > \widetilde{v}$ is at least $\geq \frac{\widetilde{v}}{\eta(\widetilde{v})}$, then for all v in this range, we have $v - \eta(v)j(v) < 0$. Thus, a man's payoff from battering, whether he accepts or rejects the offered plea bargain, is negative; he will choose $a = NB$ which gives him a payoff of zero. **Q.E.D.**

Proof of Lemma 5. Recall, at $\widetilde{v}$, $C(v) = \gamma\left[r_f + (1-q+c)w_2(v)\right]$. We have $C'(\widetilde{v}) > \gamma(1-q+c)w_2'(\widetilde{v})$. Then, since $C'' > 0$ and $w_2''(v) \leq 0$, we have $C(v) > \gamma\left[r_f + (1-q+c)w_2(v)\right]$ $\forall v > \widetilde{v}$. Thus, it is never worthwhile for malicious women to fake evidence of violence in excess of $\widetilde{v}$, and hence Proposition 7 goes through. **Q.E.D.**

Proof of Lemma 6. Suppose that a prosecutor receives a police report of violence v. The report could have been the result of a true battery or fake evidence. Then, the prosecutor's belief $\mu(v)$ that the report indicates guilt is given by

$$\mu(v) = \frac{(1-\alpha)g(v)\eta(v)}{\alpha s\gamma + (1-\alpha)g(v)\eta(v)}. \tag{17}$$

Consider the first term in the denominator of the above fraction. This shows that, subject to the level of v being compatible with false accusation and evidence faking, the probability of this false case entering the system is γ. The second term shows that if v results from true battery, the probability of the case entering the system is $\eta(v)$. Note that given $\eta(v)g(v)$ increases in v, $\mu(v)$ is increasing in v.

$\lambda(v)$, the probability that a guilty defendant charged with v rejects the plea bargain $p_L(v)$, is also increasing in v; this can easily be seen (as in the proof of Lemma 4) from the fact that a harsher plea offer prompts more rejection, and that plea offers become harsher when the charge of violence is higher.

Differentiating (6) w.r.t. v, we obtain, after simplification,

$$\frac{d\mu'}{dv} = (1-\mu)\frac{d(\lambda\mu)}{dv} > 0,$$

since $\lambda(v)$ and $\mu(v)$ have been shown to increase in v. Thus, the jury's posterior belief of guilt among those entering a trial is also higher for higher charges. Given this, it follows from Proposition 1 that both σ and p also go up when v is higher, as both increase in μ'. **Q.E.D.**

Proof of Lemma 7. Differentiating (9) with respect to v, we obtain

$$\pi'(v) = -\gamma p'(v)w_2(v) + \gamma(1-p(v))w_2'(v) - C'(v). \tag{18}$$

From Lemma 6, we know that $p'(v) > 0$. We also know that (by assumption) $w_2'(v)$ and

$C'(v)$ are both positive. Differentiating (18) once more with respect to v, we obtain

$$\pi''(v) = -\gamma p''(v)w_2(v) - 2\gamma p'(v)w_2'(v) + \gamma(1 - p(v))w_2''(v) - C''(v). \tag{19}$$

Given the parametric restrictions in the Lemma statement, and $C''(v) > 0$, (19) is negative. Thus, (9) has a unique maximizer v^*, the solution to $\pi'(v^*) = 0$. **Q.E.D.**

Proof of Proposition 8. (i) Suppose first that $p_L(v^*) = p(v^*)j(v^*) < \frac{v^*}{\eta(v^*)}$. Then, a semi-separating equilibrium exists (see Proposition 3) such that a guilty man is indifferent between accepting $p_L(v^*)$ and rejecting it and going to jury trial. The expected payoff for each such violent man from battery (at level v^*) is $v^* - \eta(v^*)p_L(v^*) > 0$, which is greater than his payoff for not battering, zero. Thus, even though reporting takes place (as $v^* > \underline{v}$), battery at v^* is not deterred, and therefore by Lemma 2, it is also safe to file fake reports of v^*. On the other hand, if $p_L(v^*) \geq \frac{v^*}{\eta(v^*)}$, and $j(v^*) \geq \frac{v^*}{\eta(v^*)p(v^*)}$, we have $v^* - \eta(v^*)p_L(v^*) \leq 0$, so violent men do not batter. By Lemma 2, this also rules out fake evidence of v^*.

(ii) Since v^* may be faked, $j(v^*)$ must be even higher than the plea offer given $p(v^*) < 1$. Hence, plea discounts are necessary.

(iii) In the proof of part (ii) of Proposition 7, we have shown that $\frac{v}{\eta(v)}$ is decreasing in v; therefore, $\frac{v^*}{\eta(v^*)}$ decreases in v^*. Thus, deterrence is easier to achieve if v^* is high.

(iv) If $v^* < \underline{v}$, then battered women do not report v^*. In this case, the perpetrators are not deterred, but no true cases are reported either; see the text just before the statement of the proposition. But then by Lemma 2, false reports do not enter the system either. **Q.E.D.**

References

A. Aizer (2010). The Gender Wage Gap and Domestic Violence. *American Economic Review*, 100, 1847–1859.

A. Aizer and P. Dal Bó (2009). Love, Hate and Murder: Commitment Devices in Violent Relationships. *Journal of Public Economics*, 93, 412–428.

D. Anderberg, H. Rainer, J. Wadsworth and T. Wilson (2016). Unemployment and Domestic Violence: Theory and Evidence. *Economic Journal*, 126, Issue 597, 1947–1979.

S. Anderson and G. Genicot (2015). Suicide and Property Rights in India. *Journal of Development Economics*, 114, May issue, 64–78.

J. Andreoni (1991). Reasonable Doubt and the Optimal Magnitude of Fines: Should the Penalty Fit the Crime? *RAND Journal of Economics*, 22(3), 385–395.

E. Arenas-Arroyo, D. Fernandez-Kranz and N. Nollenberger (2021). Intimate Partner Violence under Forced Cohabitation and Economic Stress: Evidence from the COVID-19 Pandemic. *Journal of Public Economics*, 194, February issue, 104350.

H. Avieli (2022). False Allegations of Domestic Violence: A Qualitative Analysis of Ex-Partners' Narratives. *Journal of Family Violence*, 37, 1391–1403.

S. Baker and C. Mezzetti (2001). Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial. *Journal of Law, Economics, and Organization*, 17, Issue 1, 149–167.

G.S. Becker (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76(2), 169–217.

G.S. Becker (1993). Nobel Lecture: The Economic Way of Looking at Behavior. *Journal of Political Economy*, 101(3), 385–409.

N. Berg and J-Y. Kim (2018). Plea Bargaining with Multiple Defendants and its Deterrence Effect. *International Review of Law and Economics*, 55, September issue, 58–70.

R.A. Berk, A. Campbell, R. Klap and B. Western (1992). The Deterrent Effect of Arrest in Incidents of Domestic Violence: a Bayesian Analysis of four Field Experiments. *American Sociology Review*, 698–708.

D. Bjerk (2007). Guilt Shall Not Escape or Innocence Suffer? The Limits of Plea Bargaining When Defendant Guilt is Uncertain. *American Law and Economics Review*, 9, Issue 2, 305–329.

D. Bjerk (2021). Socially Optimal Plea Bargaining with Costly Trials and Bayesian Juries. *Economic Enquiry*, 59, Issue 1, 263–279.

F. Bloch and V. Rao (2002). Terror as a Bargaining Instrument: A Case Study of Dowry Violence in Rural India. *American Economic Review*, 92, 1029–1043.

B.H. Bornstein and E. Greene (2011). Jury Decision Making: Implications for and from Psychology. *Current Directions in Psychological Science*, 20, Issue 1, 63–67.

Y-M. Chin and S. Cunningham (2019). Revisiting the Effect of Warrantless Domestic Violence Arrest Laws on Intimate Partner Homicides. *Journal of Public Economics*, 179, 1–17.

I.-K. Cho and D.M. Kreps (1987). Signaling Games and Stable Equilibria. *Quarterly Journal of Economics*, 102(2), 179–222.

B.P. Foster (2007). Analysis of Domestic Violence Costs in West Virginia and the Potential Cost of False or Unnecessary Claims. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1015102.

R.M. Gold (2011). Promoting Democracy in Prosecution. *Washington Law Review*, 86, 69–124. Available at: https://digitalcommons.law.uw.edu/wlr/vol86/iss1/8.

G.M. Grossman and M.L. Katz (1983). Plea Bargaining and Social Welfare. *American Economic Review*, 73, No. 4, 749–757.

B. Guha (2018). Secret Ballots and Costly Information Gathering: the Jury Size Problem Revisited. *International Review of Law and Economics*, 54, 58–67.

B. Guha (2020). Pretrial Beliefs and Verdict Accuracy: Costly Juror Effort and Free Riding. *B. E. J. Theor. Econ.*, 20 (2), 20180020.

B. Guha (2022). Ambiguity Aversion, Group Size and Deliberation: Costly Information and Decision Accuracy. *Journal of Economic Behavior and Organization*, 201, 115–133.

B. Guha (2023). Accomplice Plea Bargains in the Presence of Costly Juror Effort. *Games and Economic Behavior*, 142, 209–225.

B. Guha (2024). Plea Bargaining when Juror Effort is Costly. *Economic Theory*, 78, 945–977.

R. Iyengar (2009). Does the Certainty of Arrest Reduce Domestic Violence? Evidence from Mandatory and Recommended Arrest Laws. *Journal of Public Economics*, 93(1-2), 85–98.

F.X. Lee and W. Suen (2020). Credibility of Crime Allegations. *American Economic Journal: Microeconomics*, 12(1), 220–259.

S.M. Lee (2014). Plea Bargaining: On the Selection of Jury Trials. *Economic Theory*, 57, 59–88.

S. Mongrain and J. Roberts (2009). Plea Bargaining with Budgetary Constraints. *International Review of Law and Economics*, 29(1), 8–12.

K. Mukhopadhaya (2003). Jury Size and the Free Rider Problem. *Journal of Law, Economics, and Organization*, 19, Issue 1, 24–44.

T. Palfrey, H. Rosenthal and N. Roy (2017). How Cheap Talk Enhances Efficiency in Threshold Public Goods Games. *Games and Economic Behavior*, 101, January issue, 234–259.

J. Reinganum (1988). Plea Bargaining and Prosecutorial Discretion. *American Economic Review*, 78, 713–728.

N.M. Rutledge (2009). Turning a Blind Eye: Perjury in Domestic Violence Cases. *New Mexico Law Review*, 39, 149–194.

L.W. Sherman, D.A. Smith, J.D. Schmidt and D.P. Rogan (1992). Crime, Punishment, and Stake in Conformity: Legal and Informal Control of Domestic Violence. *American Sociology Review*, 680–690.

R. Siegel and B. Strulovici (2023). Judicial Mechanism Design. *American Economic Journal: Microeconomics*, 15(3), 243–270.

R. Siegel and B. Strulovici (2025). Splitting the 'Guilty' Verdict: An Analysis of a Three-Verdict System. https://sites.google.com/site/ronsiegel/.

H.V. Tauchen, A.D. Witte and S.K. Long (1991). Domestic Violence: A Nonrandom Affair. *International Economic Review*, 32, 491–511.

J.I. Turner (2013). Plea Bargaining, in *International Criminal Procedure*, 34–65 (Edward Elgar Publishing).

A. Tur-Prats (2021). Unemployment and Intimate Partner Violence: A Cultural Approach. *Journal of Economic Behavior and Organization*, 185, 27–49.

A.M. Zelcer (2014). Battling Domestic Violence: Replacing Mandatory Arrest Laws with a Trifecta of Preferential Arrest, Officer Education, and Batterer Treatment Programs. *American Criminal Law Review*, 51, Issue 2, 541–562.

A.M. Zeoli, A. Norris and H. Brenner (2011). A Summary and Analysis of Warrantless Arrest Statutes for Domestic Violence in the United States. *Journal of Interpersonal Violence*, 26, Issue 14, 2811–2833.

Supplementary Appendix (not for publication)

A.1 Additional results for Sections 4 & 5: Deterring DV under mandatory arrest

Can the prosecutor set the plea bargain in such a way that the equilibrium of Proposition 5 does not arise? In Proposition 9 we examine this possibility, showing that if the plea bargain is set such that $p_L > v$, and subject to Assumption 3, we can support a no-battering, no false-reporting equilibrium. Clearly, the equilibrium of Proposition 5 and the equilibrium of Proposition 9 cannot coexist, and which one obtains will depend on the prosecutor's choices.

Assumption 6 (Large penalty) *Suppose that in the event the jury returns a verdict of false accusation, the judge can slap the plaintiff with a penalty of* $F > \frac{r_f}{(v/j)} + (\frac{j}{v} - 1)w_2$.

Assumption 6 is stronger than Assumption 2, and thus replaces the latter.

Proposition 9 (Full deterrence–fixed harm) *Suppose Assumption 1, and Assumptions 3–6 hold. Moreover, suppose that the plea offer by the prosecutor satisfies* $\max\{v, w_1\} < p_L \leq p(0.5) \times j$. *Then a no-battering, no false-reporting* PBE *exists with the following off-equilibrium belief: upon report of any violence and rejection of plea bargain* p_L *(within the specified range), the juror's belief* $\mu' = p^{-1}(p_L/j)$ *(see Eq. (11)) about the defendant's guilt is such that the probability of reaching a correct verdict (with jurors choosing* $\sigma > 0$ *satisfying Eq. (3)) is*

$$p(\mu') = p_L/j. \tag{A.20}$$

In particular, the equilibrium can be characterized as follows:

(i) In stage 4, which is off the equilibrium path, jurors choose a verdict based on majority signals, convicting an innocent defendant incorrectly with probability $1 - p(\mu') = 1 - p_L/j$ *and acquitting him correctly with probability* $p(\mu') = p_L/j$. *Similarly, a guilty defendant will be correctly convicted with probability* $p(\mu') = p_L/j$ *and acquitted incorrectly with probability* $1 - p(\mu') = 1 - p_L/j$. *The jury will acquit the defendant if no one observes a signal. When jurors acquit the defendant (with or without signals), they also blame the plaintiff for a false accusation.*

(ii) At stage 3, innocent men will reject the plea offer with probability $\widetilde{\lambda}_I = 1$ *and guilty men reject it with some arbitrary probability* $0 < \widetilde{\lambda}_G < 1$.

(iii) At stage 2, women choose $\chi_B = R$ *and* $\chi_{NB} = N$;

(iv) At stage 1, violent men choose $a = NB$.

Proof of Proposition 9. (i) Given the posited equilibrium of no battering and no false reporting, a report of DV triggers an off-equilibrium continuation game. Suppose that the prosecutor makes a plea offer $v < p_L \leq p(0.5) \times j$. Now, given the belief μ' (as stated) and the positive σ induced in stage 4 as a result, the jury verdict based on majority signals is clearly sequentially rational; if all fail to observe a signal, choosing to acquit is optimal because the default belief $\mu' = p^{-1}(p_L/j) \leq 0.5$. Finally, the jury decision to blame the plaintiff for false accusation when acquitting the defendant, which is essentially a reversal of belief from $\mu' > 0$ to $\mu' = 0$, is consistent with the definition of PBE because the continuation game is still an off-equilibrium play and any signals (or the lack of thereof) failed to validate $\mu' > 0$ with sufficient confidence.[26] The stated probabilities of conviction and acquittal, correctly or incorrectly, follow from the construction of $p(\mu')$ in Eq. (A.20).

(ii) By Proposition 3, the response of the defendant to the plea offer must be semi-separating. A guilty man will receive a residual payoff of $-p_L$ by accepting the plea offer, and his expected payoff from rejecting the plea offer and proceeding to trial is $-p \times j$. By (A.20), $-p_L = -p \times j$, thus a guilty man must be indifferent between accepting and rejecting the plea offer. We let him choose an arbitrary probability $0 < \widetilde{\lambda}_G < 1$ of rejecting the plea offer. Since we can choose $\widetilde{\lambda}_G$ arbitrarily close to 1, by applying Lemma 1 we can conclude that if the defendant is innocent, then he will surely reject the plea offer: $\widetilde{\lambda}_I = 1$.

(iii) First, suppose that a woman has been battered, which is a deviation by her partner. If she reports, the accused will be convicted with probability $p(\mu') = p_L/j$, as shown in part (i). In that case, the woman will receive an expected payoff of $(1 - \widetilde{\lambda}_G)(r + w_1) + \widetilde{\lambda}_G\big[(p_L/j)(r - m + w_2) + (1 - (p_L/j))(r - m - w_3 - F)\big]$.[27] If she does not report, her payoff is $-h$. See Fig. 2. Given that $r - m - w_3 - F > -h$ (Assumption 3), $r - m + w_2 > r - m - w_3 - F$ and $r + w_1 > 0$, clearly the payoff from reporting dominates the payoff from not reporting. Hence she will report.

Next, suppose that a woman is not battered. If she is not malicious but deviates and reports, the innocent man will always go to trial and will be mistakenly convicted with probability $1 - p(\mu') = 1 - p_L/j$, as shown in part (i). This yields her a payoff of $(1 - p_L/j)(-\mathpzc{r}' + w_2) + (p_L/j)(-\mathpzc{r}'' - F)$. If she does not report, her payoff is 0. See Fig. 2. Applying Assumption 4, it follows that a non-malicious woman would not report falsely. On the other hand, if the woman is malicious, by reporting she collects a payoff of $(1 - p_L/j)(r_f + w_2) + (p_L/j)(r_f - F)$; instead, if she does not report her payoff is 0. The

[26] We are exploiting zero restriction on off-equilibrium beliefs built into the definition of PBE.

[27] Recall, acquittal by the jury comes with the additional assertion that the reporting by the plaintiff is malicious (i.e., a false accusation). Therefore, the plaintiff pays the penalty F.

payoff from not reporting will dominate the payoff from reporting if

$$(1 - p_L/j)(r_f + w_2) + (p_L/j)(r_f - F) < 0$$
$$\text{i.e.,}\quad F > \frac{r_f}{(p_L/j)} + (\frac{j}{p_L} - 1)w_2,$$

which will be satisfied given Assumption 6; note that $\frac{v}{j} < \frac{p_L}{j} \leq 1$ by the specified range of p_L. Thus, the malicious woman will not report falsely either.

(iv) Finally, consider the incentives of a violent men. By part (iii) result, if such a man chooses $a = B$, his woman will choose $\chi_B = R$, i.e., lodge a report of violence. Then his expected payoff will be $v - p_L < 0$ (note that even if he rejects the plea offer, his expected punishment from trial will also be equal to $p_L = p(p_L) \times j$), while his payoff from choosing $a = NB$ is 0. Thus, he prefers not to batter. **Q.E.D.**

Note that for the full-deterrence result we had to impose a stricter penalty F for false accusation than we had assumed for Proposition 5. This is to deter a non-malicious woman's deviation to reporting when she is not battered. Talking about a non-malicious woman reporting may sound like an oxymoron. As one can see in Fig. 2, the distinguishing characteristic of a non-malicious woman from that of a malicious woman is that the former receives a negative payoff from reporting when not battered, whereas the latter enjoys reporting ($-\mathscr{r}$, $-\mathscr{r}'$ and $-\mathscr{r}''$ vs. r_f). The reason we still need to consider the incentive for a non-malicious woman to do false reporting is because by doing so she earns positive compensation w_1 and w_2 when the defendant pleads guilty or is convicted by the jury.

Thus, provided some parametric restrictions are satisfied, the prosecutor can set a sufficiently harsh plea bargain to deter actual incidents of DV and malicious reporting.

Combining Propositions 5 and 9, we can summarize the relationship between plea bargain punishment and deterrence outcome (i.e., fraction of men against whom their women have reported violence, truthfully or maliciously) in Fig. 10.

Proposition 10 (Deterrence—heterogeneous harm) *Suppose Assumptions 1 and 4 hold and suppose that we have*

$$\frac{r_m + w_2(1 - \frac{\bar{v}}{j})}{(\bar{v}/j)} < F < \frac{r - m - w_3 + (\frac{\bar{v}}{j})(w_2 + w_3)}{(1 - \frac{\bar{v}}{j})}.$$

Moreover, suppose that the plea offer by the prosecutor satisfies

$$\bar{v} \leq p_L \leq p(0.5) \times j.$$

Then a no-battering, no false-reporting equilibrium exists with the following off-equilibrium belief: upon reporting of any violence and rejection of plea bargain p_L *(within the speci-*

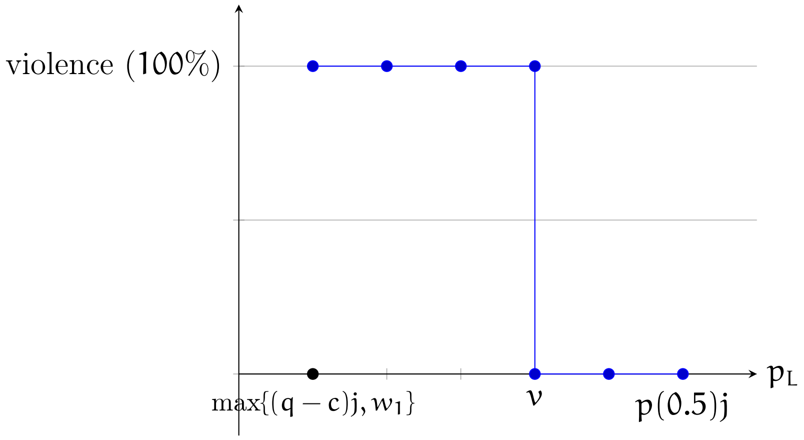


Figure 10: %reported violence against plea bargain punishment p_L. Proposition 5 (full violence) in the range $\max\{(q-c)j, w_1\} \le p_L \le v$, and Proposition 9 (complete deterrence) when $\max\{v, w_1\} < p_L \le p(0.5) \times j$.

fied range), the juror's belief $\mu' = p^{-1}\left(\frac{p_L}{j}\right)$ (see Eq. (11)) is such that the probability of reaching a correct verdict (with jurors choosing $\sigma > 0$ satisfying Eq. (3)) is $p(\mu') = \frac{p_L}{j}$. In particular, the equilibrium can be characterized as follows:

(i) In stage 4, which is off the equilibrium path, jurors choose a verdict based on majority signals, convicting an innocent defendant incorrectly with probability $1-p(\mu') = 1-\frac{p_L}{j}$ and acquitting him correctly with probability $p(\mu') = \frac{p_L}{j}$. Similarly, a guilty defendant will be correctly convicted with probability $p(\mu') = \frac{p_L}{j}$ and acquitted incorrectly with probability $1-p(\mu') = 1-\frac{p_L}{j}$. The jury will acquit the defendant if no one observes a signal. When jurors acquit the defendant (with or without signals), they also blame the plaintiff for a false accusation.

(ii) At stage 3, innocent men will reject the plea offer with probability one, and guilty men reject it with some arbitrary probability $0 < \widetilde{\lambda}_G < 1$.

(iii) At stage 2, women choose $\chi_B = R$ and $\chi_{NB} = N$.

(iv) At stage 1, violent men choose $a = NB$.

Proof of Proposition 10. The proof of steps (i) and (ii) is identical to the proof of the corresponding parts of Proposition 9. Now, consider part (iii). We are given $F < \frac{r-m-w_3+(\bar{v}/j)(w_2+w_3)}{1-(\bar{v}/j)}$. Rearranging terms, this gives us

$$\frac{\bar{v}}{j}(r-m+w_2) + (1-\frac{\bar{v}}{j})(r-m-w_3-F) > 0. \tag{A.21}$$

Since $p = p_L/j$ and $p_L \ge \bar{v}$, note that p is at least $\bar{v}/j$. Therefore, since $w_2 > -w_3 - F$, (8) implies

$$p(r-m+w_2) + (1-p)(r-m-w_3-F) > 0. \tag{A.22}$$

Now the LHS of (A.22) is the payoff that a battered woman receives if she reports *and* the case goes to trial, in which case she secures a conviction with probability p, but is accused of malicious intent with complementary probability. The actual payoff for the report is even higher than this, as if the plea bargain is accepted, she gets a strictly positive payoff $r + w_1$. Eq. (A.23) shows that the payoff from the reporting is also always positive:

$$\widetilde{\lambda}_G \left[p(r - m + w_2) + (1 - p)(r - m - w_3 - F) \right] + (1 - \widetilde{\lambda}_G)(r + w_1) > 0 > -v. \quad \text{(A.23)}$$

Thus, it is always more than her payoff from not reporting, which is $-v$ for a woman experiencing violence v (by definition any $-v$ is negative). So we have $\chi_B = R$. The proof that non-malicious women do not make false reports is identical to the corresponding proof in Proposition 5 (we invoke Assumption 4). Turning to the incentives for malicious women, we are given $\frac{r_m + w_2(1 - \bar{v}/j)}{(\bar{v}/j)} < F$. Rearranging terms, we get

$$\frac{\bar{v}}{j}(r_m - F) + (1 - \frac{\bar{v}}{j})(r_m + w_2) < 0. \quad \text{(A.24)}$$

Given that p is at least $\bar{v}/j$, this implies (since $rm < F$) that

$$p(r_m - F) + (1 - p)(r_m + w_2) < 0. \quad \text{(A.25)}$$

The LHS of (A.25) shows a malicious woman's payoff from making a false report; her innocent man rejects any plea offer with probability one, so a trial occurs, where the woman is awarded a compensation w_2 in the event of an incorrect conviction, while otherwise she is accused of malicious intent. Her payoff from making a false report is thus always smaller than her payoff from not lodging such a report. Accordingly, $\chi_{NB} = N$.

Turning to part (iv), any potentially violent man knows that any act of violence will be reported. Since his payoff from not battering, zero, exceeds the negative payoff $v - p_L$ (the payoff is negative given $v < \bar{v} \leq p_L$), he chooses $a = NB$ regardless of the realization of v. **Q.E.D.**

The intuition behind why F should be neither too large nor too small is as follows. In the off-equilibrium event that men batter, their women must be prepared to report them (to sustain a no-battery outcome in equilibrium). However, there is a small chance that if a woman reports and the case goes to trial, the man may be wrongly acquitted and then the woman is slapped with a fine. The fine must not be so large that it deters true reporting. But if the fine is too small, it encourages false reporting.

Note that the lower limit on F decreases in $\bar{v}$, while its upper limit increases in $\bar{v}$. Therefore, a high $\bar{v}$ increases the likelihood that the lower limit for F is smaller than the upper limit, so that there is a non-empty parameter range where true reporting is encouraged and false reporting is discouraged. Intuitively, a high $\bar{v}$ also implies a harsh plea bargain and a high p. Now, false reporting is deterred for a relatively modest fine as the chances of a correct acquittal and being charged with malicious intent are higher for a false report; accordingly, the lower limit for F is lowered. At the same time, a high p means that there is very little chance of an incorrect acquittal in the case of true violence. Thus, battered women feel that there is very little chance of being slapped with a fine if they report, hence they are willing to report even for a rather high F. Thus, the upper limit for F is raised.

The second feature that makes the no-battery, no false-reporting equilibrium more feasible is if $p(0.5) \times j$ is likely to be high, as it needs to be higher than $\bar{v}$. This can happen if $p(0.5)$ is high for a given μ' (this can happen, for example, if c is low or if n is chosen appropriately).

A.2 Effects of beliefs (μ') and jury size (n) on jury attention and verdict accuracy

In Proposition 1 we had established analytically the comparative static effects of beliefs μ' on jury efforts and the verdict accuracy. We did not study the effect of jury size n. In the following, we do some numerical computations to further illustrate the comparative static effects and demonstrate the effect of jury size n.

Table A.2: σ for $c = 0.1, q = 0.9$

	$\mu' = 0.5$	$\mu' = 0.4$	$\mu' = 0.3$
$n = 3$	0.56	0.456	0.304
$n = 5$	0.2931	0.2402	0.1591
$n = 7$	0.2063	0.1673	0.1091
$n = 9$	0.1591	0.1283	0.0829
$n = 11$	0.1294	0.104	0.0669
$n = 13$	0.1091	0.0874	0.0561
$n = 15$	0.094	0.0754	0.0483

Tables A.2 to A.5 show values of σ and p for different values of μ' and n. Tables A.2 and A.3 derive σ and p respectively, for the case where $q = 0.9, c = 0.1$. Tables A.4 and A.5 do so for a more precise signal and a lower cost of attention, $q = 0.99, c = 0.02$. We illustrate numerically, as argued in Proposition 1, that fixing q, c, and n, both σ and p increase as μ' increases (we consider μ' less than or equal to 0.5, so that acquittal is the default option). We have restricted to the domain where $\mu' > 1 - q + c$, so that jurors always choose to pay attention with positive probability.

Intuitively, when $\mu' = 0.5$, jurors are very uncertain about guilt or innocence, and hence have a greater tendency to pay attention to signals – an instinct which decreases

Table A.3: p for $c = 0.1, q = 0.9$

	$\mu' = 0.5$	$\mu' = 0.4$	$\mu' = 0.3$
$n = 3$	0.8785	0.8645	0.8415
$n = 5$	0.8398	0.8354	0.8235
$n = 7$	0.8302	0.8273	0.8182
$n = 9$	0.8276	0.8232	0.8151
$n = 11$	0.8238	0.8215	0.8137
$n = 13$	0.8215	0.8206	0.8114
$n = 15$	0.8212	0.8188	0.8106

Table A.4: σ for $c = 0.02, q = 0.99$

	$\mu' = 0.5$	$\mu' = 0.4$	$\mu' = 0.3$
$n = 3$	0.8357	0.7653	0.6994
$n = 5$	0.57271	0.4919	0.4376
$n = 7$	0.43	0.3632	0.3187
$n = 9$	0.34	0.2871	0.2501
$n = 11$	0.29	0.2372	0.2056
$n = 13$	0.25	0.202	0.1746
$n = 15$	0.22	0.1758	0.1517

Table A.5: p for $c = 0.02, q = 0.99$

	$\mu' = 0.5$	$\mu' = 0.4$	$\mu' = 0.3$
$n = 3$	0.9934	0.9898	0.9868
$n = 5$	0.9895	0.9816	0.9782
$n = 7$	0.9873	0.9781	0.9747
$n = 9$	0.9856	0.9768	0.9724
$n = 11$	0.9867	0.9755	0.9719
$n = 13$	0.9872	0.9734	0.9725
$n = 15$	0.988	0.9738	0.9704

when their updated beliefs become more skewed in any direction (in this case towards innocence) as μ' deviates from 0.5. Thus, jurors "free ride" on their updated beliefs as these beliefs become more extreme. This carries over to p – the lower probability of jurors being attentive with a fall in μ' translates into less accurate verdicts overwhelming the effect of uninformed decisions being correct more often (due to more sharply skewed beliefs, which are correct more often). (This is shown in Guha (2020) for the case of correlated signals.)

Fixing q, c and μ', σ always decreases in n, jury size, as a larger number of jurors induces each juror to reduce the probability of paying attention; the chance of being pivotal shrinks. The effect of n on p need not necessarily be monotonic, however. The effect of less attention being paid in bigger panels tends to pull down p as panels become larger. However, a counteracting effect of more jurors receiving (conditionally) independent signals and pooling information tends to pull up p. In Table A.3, p decreases in n holding μ' fixed. Thus, when q is not extremely large, and c not very small, the free-riding effect prevails, so that in Table A.3 p decreases in n; this is reminiscent of a similar observation in Mukhopadhaya (2003). However, in Table A.5, which is for an extremely precise signal ($q = 0.99$) and a very low cost of juror attention ($c = 0.02$), we notice a non-monotonicity for all three values of μ'. Intuitively, free riding here does not have as bad an effect on p as in Table A.3, because the signal is very precise, and σ is relatively high, despite free riding, owing to c being so low. When $\mu' = 0.5$, p falls in n until n reaches 9, and starts increasing beyond this as the second effect, of pooling independent signals, overpowers the effect of greater free riding in a larger panel. When $\mu' = 0.3$, p falls in n until n reaches 11, increases, and then decreases again, so non-monotonicity exhibits an irregular pattern. When $\mu' = 0.4$, p falls in n until n reaches 13, and increases afterwards. The relatively muted effect on p's increase is because even though the probability of reaching ties starts to increase, it does not lead to correct verdicts often enough (as the belief is not extreme enough) and this is not sufficient to overpower the effect of lower attention, until jury size becomes very high. $\|$

Supplementary mathematica derivation in support of Prop 1: Checking the sign of the derivative of sigma:

```
In[1]:= Element[j, Integers]
```

Out[1]= $j \in \mathbb{Z}$

```
In[2]:= Element[n, Integers]
```

Out[2]= $n \in \mathbb{Z}$

The following is the derivative of the LHS of Eq. (3) w.r.t. sigma. We show that the sign is not necessarily negative for the entire range of parameters. In particular, the derivative plot below changes from negative to positive when sigma is very high.

```
In[3]:= der[q_, σ_, n_] :=
          -(1 - σ)^{n - 3} *
            (Sum[((n - 1)!) /
                    ((n - 1 - 2*j)! * j! *
                       j!) * (q * (1 - q))^j *
                  (σ / (1 - σ))^{2*j - 1} *
                  ((n - 1) σ - 2*j),
                {j, 1, (n - 1) / 2}] +
               (n - 1) * (1 - σ))
```

In[4]:=
```
der[q, σ, n]
```

Out[4]=

$$\left\{-(1-\sigma)^{-3+n}\left((-1+n)\ (1-\sigma)+\frac{-1+n}{-1+\sigma}-\frac{2\ (-1+n)\ \sigma}{-1+\sigma}+\frac{(-1+n)\ \sigma^2}{-1+\sigma}-\frac{(-1+n)\ \text{Hypergeometric2F1}\left[\frac{1}{2}-\frac{n}{2},1-\frac{n}{2},1,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}+\frac{2\ (-1+n)\ \sigma\ \text{Hypergeometric2F1}\left[\frac{1}{2}-\frac{n}{2},1-\frac{n}{2},1,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}-\frac{(-1+n)\ \sigma^2\ \text{Hypergeometric2F1}\left[\frac{1}{2}-\frac{n}{2},1-\frac{n}{2},1,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}-\frac{4\ (-1+n)\ q\ \sigma\ \text{Hypergeometric2F1}\left[\frac{3}{2}-\frac{n}{2},2-\frac{n}{2},2,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}+\frac{2\ (-1+n)\ n\ q\ \sigma\ \text{Hypergeometric2F1}\left[\frac{3}{2}-\frac{n}{2},2-\frac{n}{2},2,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}+\frac{4\ (-1+n)\ q^2\ \sigma\ \text{Hypergeometric2F1}\left[\frac{3}{2}-\frac{n}{2},2-\frac{n}{2},2,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}-\frac{2\ (-1+n)\ n\ q^2\ \sigma\ \text{Hypergeometric2F1}\left[\frac{3}{2}-\frac{n}{2},2-\frac{n}{2},2,-\frac{4\ (-1+q)\ q\ \sigma^2}{(-1+\sigma)^2}\right]}{-1+\sigma}\right)\right\}$$

In[5]:=
```
der[0.7, 0.3, 5]
```

Out[5]=
```
{-0.920063}
```

In[6]:=
```
der[0.7, 0.4, 5]
```

Out[6]=
```
{-0.554342}
```

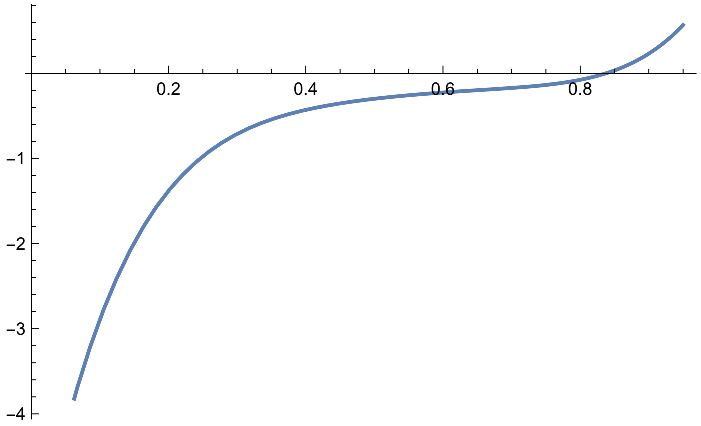
In[7]:=
```
der[0.7, 0.9, 5]
```

Out[7]=
```
{0.404694}
```

In[8]:=
```
der[0.7, 0.4, 7]
```

Out[8]=
```
{-0.429053}
```

```
In[9]:= der[0.7, 0.5, 7]

Out[9]= {-0.298459}

In[10]:= der[0.7, 0.6, 7]

Out[10]=
        {-0.224927}

In[11]:= der[0.7, 0.7, 7]

Out[11]=
        {-0.171436}

In[12]:= der[0.7, 0.8, 7]

Out[12]=
        {-0.0757991}

In[13]:= der[0.7, 0.9, 7]

Out[13]=
        {0.231809}

In[14]:= Plot[der[0.7, σ, 7],
          {σ, 0.01, 0.95}]

Out[14]=
```


0.2
0.4
0.6
0.8
-1
-2
-3
-4