American Philosophical Quarterly
Volume 60, Number 2, April 2023

INTRODUCTION, SUMMARY, QUESTIONS FOR THE FUTURE

Philosophical Dimensions of the Trial

Lewis Ross, Miguel Egler, and Lisa Bastian

Legal trials are a rich source of philosophical interest. Familiar moral and political topics often discussed in the abstract by philosophers find concrete expression in the high-stakes environment of the courtroom. Dusty epistemological questions about evidence and testimony become problems of momentous importance in the law, where errors can lead to the imprisonment of the innocent. It is therefore no surprise that legal philosophy is currently an exciting and fruitful area of research. This special issue of the *American Philosophical Quarterly* showcases some of this research, focusing on philosophical questions raised by the legal trial.

This special issue brings together a collection of original papers by: (i) Zachary Hoskins (Nottingham), (ii) Georgi Gardiner (Tennessee), (iii) Talia Fisher (Tel Aviv), (iv) Joe Slater (Glasgow), (v) Jeremy Davis & Duncan Purves (Florida), (vi) Christoph Winter, Nicholas Hollman & David Manheim (various). The topics discussed are of wide social importance, broadly unified by a concern with how societies should administer justice in the future. Legal systems around the world are grappling with new technologies in the courtroom (e.g., algorithmic decision-making tools for predicting recidivism) and reconsidering old technologies (e.g., using a jury of lay peers to discern true from false), they are facing difficult ethical criticisms (e.g., the low conviction rates for sexual criminality) and confronting new paradigms for justice (e.g., the customization of trials through contractual agreement). Within these debates, there is ample room for philosophers to put their moral, political, and epistemological sophistication to good social work.

Below we introduce the papers in this special issue and flag questions for further discussion raised by these contributions.

1. Collateral Legal Consequences and Criminal Sentencing, Zachary Hoskins

Punishment in the form of "hard treatment"—paradigmatically, imprisonment—is among the most heavily discussed topics in legal philosophy. But there are a host of consequences that flow from criminal conviction which are distinct from the formal imposition of hard treatment. For example, convicted persons can face restrictions in their ability to vote, to secure housing or employment, or the publication of their criminal record. These "collateral consequences" are the topic of Zachary Hoskins' paper.

Hoskins' main question is whether the onerous nature of these collateral legal consequences (CLCs) should be taken into account by the judge when issuing a sentence. For instance, should a judge moderate a sentence in anticipation of the onerous nature of various CLCs associated with conviction? One way to approach this question is to consider whether these CLCs should properly be regarded as a type of punishment. If they *are* then, arguably, CLCs should be considered when sentencing. Hoskins defends a conception of punishment on which it must be *intentionally burdensome and intended to convey censure* (the ICB account). Hoskins situates this position in relation to challenges to each conjunct of the ICB account due to Bill Wringe and Ambrose Lee. Many CLCs *are* intentionally burdensome or intended to convey censure: voting restrictions, for example, are often justified as a way of expressing the idea that the offender has alienated themselves from the wider community and is currently undeserving of being regarded as a full member of it. Hoskins argues that CLCs which do amount to punishments properly considered should factor into sentencing decisions—else they risk being a type of "invisible punishment" that are not properly accounted for when courts seek to do justice by distributing burdens appropriately.

Two key questions for the future. First, what should we think about CLCs that are nonpunitive, that is, those that do not formally count as varieties of punishment? Second, how should we consider *informal* collateral consequences that are not imposed by the legal system? The social stigma of conviction is a prime example. Hoskins suggests that such informal consequences should not be considered in the sentencing decision. However, in some cases, the burdensomeness of the informal consequences may indeed outweigh that of the formal sentence and any CLCs associated with it. So, we need to think carefully about the nature and extent of the state's responsibilities when it comes to managing the effect of these informal consequences on the convicted person.

2. Corroboration, Georgi Gardiner

Georgi Gardiner focuses on the notion of "corroborating" evidence: evidence that supports—or, we might say, confirms—a proposition which already has support from some initial evidence. How should we think about the epistemic power of corroborating evidence? Why does corroborating evidence often have such a powerful psychological effect in bringing us to endorse a proposition? These are some of the questions Gardiner takes up in her contribution.

Gardiner contends that the force of corroborating evidence cannot be captured by appealing to probabilities alone—that is, corroborating evidence is powerful *not just* because it makes a proposition more likely, given the evidence. For example, if we have a proposition that is already exceedingly likely, Gardiner suggests that appealing to the probability-raising power of corroborating evidence will underplay the psychological and epistemic import of new corroborating evidence.

Rather, Gardiner defends a "relevant alternatives" framework for thinking about corroboration. The key idea behind the relevant alternatives approaches—here, and in epistemology generally—is that epistemic standing depends on which (relevant) error possibilities can be ruled out. Corroborating evidence is powerful because it rules out previously unexcluded error possibilities and serves to guide further inquiry. To see what is meant here, consider the following. As Gardiner points out, a body of evidence can both be superficially very strong (seeming to make a proposition very likely) while simultaneously being weak (because it doesn't say anything about whether the evidence itself is misleading). A written confession is one example: that someone confesses to a crime (often) makes it likely they are guilty, even

though a confession doesn't speak to the error possibility that the confession was coerced. Corroborating evidence here would shrink the space of uneliminated error possibilities—for example, alternatives on which the confession was coerced—and help us work out which error possibilities remain, and possibly in need of further investigation.

The abstract topic of corroboration has real normative legal bite. Gardiner entertains the idea that "it is always a relevant alternative for affirmative legal verdicts that any single piece of evidence is misleading," which would mean that, on a relevant alternatives framework, corroboration would be a *requirement* for a positive legal verdict. As Gardiner notes, a useful application of a corroboration requirement would be to the much-discussed proof paradox. However, a question for the future is whether such a requirement is attractive across the board. Some legal systems do in fact impose "corroboration requirements" on criminal trials, requiring two independent sources of evidence for conviction. The Scottish legal system, for instance, recently underwent a period of angst about whether the retention of a corroboration requirement was making it too hard to secure convictions for sexual offences. In such cases, we need to engage in the difficult balancing act of weighing the need to secure justice for serious offences against the imperative to avoid miscarriages of justice.

3. Trial by Design, Talia Fisher

What if we could customize the trial process? A standard model of the trial is one on which the procedural rules governing it are set in advance by the relevant legal system. In this sense, trial procedure is not a matter of negotiation, but a package of rules that ensures that every complaint receives "due process" by the legal system. In her paper, Talia Fisher explains and addresses the idea of "trial by design": where the contours of trial procedure are determined by agreement between those who are—or could be—subject to them. As Fisher demonstrates, legal systems are currently grappling with the possibility of trial by design and there is a real need for philosophical contributions to these debates. Illustrative examples of customization include:

- Agreements that certain types of evidence are (in)admissible
- Agreements about how to resolve certain factual disputes (e.g., use of a lie detector)
- Waiving certain procedural right: for example, to a jury, to appeal, to cross-examine witnesses
- To deviate from default rules concerning cost allocation
- To judge disputes against a different standard of proof

As Fisher points out, the idea of trial by design has several advantages, particularly in the domain of civil law. In abstract terms, it can increase the autonomy of the parties to a legal case, by allowing them to resolve their disputes in a manner of their choosing. A more concrete strength of customization concerns efficiency. This can happen when parties strike agreements that reduce expense, for instance by forgoing trial by jury. Alternatively, it could be more efficient for parties to agree in advance that certain topics will not be disputed in court. This can solve Prisoner's Dilemma-type scenarios, where the worst option is to underprepare for litigation, with the best option being a mutual agreement not to litigate on certain—minor—issues. However, despite these strengths, the idea of trial by design brings with it several concerns.

Firstly, there is a tension between the customization of trial procedure and the truth-seeking function of the trial. If parties have free reign to determine the terms of their trial, this brings with it the possibility of setting terms that are markedly less reliable than the default—for example, by excluding reliable evidence, including reliable evidence, or putting hard-to-satisfy conditions on the success of a lawsuit. To the extent that accuracy

at trial is a *public* as well as a private good, trial by design has the potential to be problematic. (Although, of course, a customized trial could be *more* exacting than the default package.)

Secondly, as Fisher hints at, there is a clear risk in customization that entrenches power differentials between the parties. It is objectionable to allow the powerful to strike a bargain that skews legal dispute resolution in their favor, making it harder for the weaker party to have their claim vindicated than it would be under the default procedures. This is especially risky where agreement to unequal adjudication bargains could be made a prerequisite for entering into various other contractual agreements with a powerful entity.

Thirdly, and relatedly, many of the most obvious advantages and applications of trial by design appear in the context of civil law. There are difficult questions about whether any customization is desirable in the context of criminal law, although there are increasingly familiar discussions about the ethics of the accused striking various bargains with the prosecution to ensure various types of leniency. As Fisher suggests, there are deep and important questions here about whether the (default) procedural protections given to the accused are essentially personal protections—and thus apt to be waived—or components of public justice.

4. Just Judge: The Jury on Trial, Joe Slater

In his contribution, Joe Slater considers the defensibility of trial by jury. The debate over whether we ought to use a lay jury—rather than a professionalized judiciary—to serve as the "fact-finder" in a criminal case is a venerable one. Many political virtues have been claimed for the jury, as a way of instantiating or symbolizing democracy, as a way of rooting criminal judgements in the conscience of the community, or even as a ward against oppressive prosecutions. On the other side, there are long-standing worries about the reliability of juries and their susceptibility to bias. However, there is a danger of conducting this debate in a monolithic way: suggesting that juries are either defensible or indefensible *tout court.* Slater advances this debate by advancing criteria that he thinks allows us to determine whether we should use a jury trial in a particular context. He advances four distinct criteria:

1. Jury Failure: juries must systematically fail to deliver the verdict warranted by the evidence
2. Social Costs: the above-mentioned failure must have serious social implications, for example, concerning recidivism or public confidence in the legal system
3. No Easy Fixes: the systematic failures must not be easy to fix within the jury trial paradigm, for example, through procedural changes
4. Judges Better: there must be evidence that judge-only trials would perform better.

Usefully, these criteria may be satisfied in some criminal trials but not in others, leaving open the possibility of a hybrid system for criminal adjudication.

Putting these criteria to work, Slater discusses the difficult case of using juries in sexual trials. Some empirical evidence, and informed speculation, suggests that the use of jury trials may be (part) responsible for the low conviction rate for sexual criminality. One prominent suggestion is that jurors are prey to a variety of "rape myths" that hinders their ability to correctly evaluate evidence in these contexts. (Of course, any discussion of empirical study of jurors is fraught: researchers are not permitted to study real jury deliberations—we must make do with either "mock jury" simulations or, less often, with impressionistic post trial surveys of real jurors.) Slater argues that in the case of sexual trials, all four of his criteria for abandoning jury trials are satisfied and that we ought to

prefer judge-only trials. There are, of course, questions regarding such a recommendation. One question is whether it might be the case, due to the nature of such crimes, that it is simply harder for investigators in sexual cases to gather evidence as strong as evidence in other types of cases, such as murder or theft. Another question is what effect a hybrid approach—for example, a jury for murder trials, but not for rape trials—would have on public confidence in the legal system and in the security of sexual convictions in particular. Finally, a fascinating broader question raised by Slater's paper is whether his criteria would disallow juries in other types of case, for example complex fraud trials, or cases where members of society display other types of bias, such as relating to socio-economic grouping or race.

5. Value Alignment for Advanced Artificial Judicial Intelligence, Christoph Winter, Nicholas Hollman, and David Manheim

The rapid development and growth of Artificial Intelligence systems (AI) has enabled them to play increasingly central roles in various aspects of our lives. This widespread use of AI introduces many challenges, as autonomous intelligent systems can often fail to operate in accordance with programmers' values. This *value alignment problem* for AI has gained added importance in recent years as some legal systems have begun to use AI to support or replace parts of judicial decision-making, for example, to decide small scale civil suits, to advise members of the public, or even make recommendations (e.g., about parole) to the judiciary. Although current AI has limited capabilities, scholars have considered the possibility of advanced artificial judicial intelligence (AAJI) that matches or surpasses human judicial decision-making. In their contribution, Christoph Winter, Nicholas Hollman, and David Manheim consider *how* AAJI can be made to incorporate the norms, ideals, and goals of the judiciary so as to safely constrain their operations and ensure that they function in accordance with our values. To develop these points, they first note that discussions about the fundamental values of the judiciary abound. Core juridical values include, among others, procedural fairness, trust in the courts, equality, fairness, and equal treatment under the law. Beyond these internal values, external constraints from the legislative and executive branches, as well as public opinion should presumably also have a normative role in defining the law. Despite such a large body of work and relative agreement on which values should guide the judiciary, the literature lacks the adequate specification needed to build value aligned AAJI. Winter, Hollman, and Manheim explain that such conflicts pose a significant challenge for adoption of AAJI, as the failure to *specify* how these systems should respond to such conflicts among values may produce (morally, politically, and legally) problematic decisions. In light of such issues, they call for more work on developing clear guidelines for how to navigate conflicts among judiciary values, so as to arrive at safe specifications for the operations of autonomous intelligence systems.

Winter, Hollmann, and Manheim then turn to considerations for how to monitor the workings of AAJI. They identify four assurance mechanisms for doing so. The first is *verifiability,* which concerns abilities to assess the extent to which the development and deployment of AAJI are aligned with judiciary values. They then discuss *transparency* and *interpretability* which refer to the ability to understand the operations of AAJI and how it reaches decisions. And lastly, they consider *interruptibility*, which is the capacity to stop or alter unwanted behaviors. As they argue, these four mechanisms for monitoring and controlling AAJI are required if we are to ensure that they operate as intended and can safely be put to use in the legal system.

6. Should Algorithms that Predict Recidivism Have Access to Race? Duncan Purves and Jeremy Davis

Duncan Purves and Jeremy Davis focus on recent debates surrounding the use of algorithmic risk assessment tools in sentencing, parole, and bail decisions. This is now a common practice in the American criminal justice system, even if one that is shrouded in controversy. Among the most notable controversies concerns influential yet contested criticisms of an algorithmic risk-assessment tool (COMPAS): some alleged that this tool was unfairly marking Black defendants as having a higher risk of recidivism. In light of this debate, many called for the abolishment of algorithmic risk-assessment tools in judicial settings, whereas others argued that we should make efforts to align the operations of such tools to our moral and judiciary values. Regarding the latter option, two main approaches have been proposed. The first is to establish distinct risk thresholds that are sensitive to defendants' racial profiles. This could mean, for example, that we make it so that the score required for labelling Black defendant as "high risk" would be higher than that required to label a White defendant as such. The second option is to create distinct racial "tracks," so that risk-assessment tools evaluate White and Black defendants differently, for example, by using different criteria for each, or weighting the same criteria in different ways for each.

Purves and Davis carry out a detailed assessment of these two approaches. They begin by examining Deborah Hellman's recent arguments for thinking that the use of different racial tracks is legally permissible, whereas implementing distinct risk thresholds is not. The purported difference is that only the latter would amount to a form of disparate treatment insofar as it gives defendants' race a *direct* causal role in determining legal decision-making, and because appeal to racial categorizations in defining risk thresholds entails a form of (impermissible) racial profiling. Purves and Davis argue that Hellman's definition of disparate treatment fails to identify a genuine distinction in the legal permissibility of the two approaches. However, they argue that there is indeed a morally salient difference between them, as the use of distinct risk thresholds would hold White defendants to harsher legal standards—making it thereby *morally* problematic. By contrast, they contend that implementing different racial tracks would set standards that are sensitive to the predictive values of each racial group, being thereby to the advantage of the average member of each one, even though some defendants in each group will fare worse. In sum, Purves and Davis argue that there are morally important distinctions between two much-discussed approaches for reforming the use of algorithmic tools. However, it remains an open question whether these moral distinctions render it (im)permissible to implement any of these two approaches all-things-considered.

Lewis Ross
Lakatos Building
London School of Economics
London, WC2A 2AE
L.Ross2@lse.ac.uk

Miguel Egler
Tilburg University
Warandelaan 2
Tilburg, The Netherlands 5037 AB
m.egler@uvt.nl

Lisa Bastian
VU Amsterdam
De Boelelaan 1105
1081HV Amsterdam
l.bastian@vu.nl